A Sign of ill Intent: Unjustified Harming Signals Harm was Intentional

A thesis submitted in partial fulfilment of the requirements for the Degree of Master of Science in Psychology in the University of Canterbury

By

Stephen James Rowe

School of Psychology, Speech and Hearing

University of Canterbury

June 28, 2021

Supervised by

Andrew Vonasch, University of Canterbury

Andrew Monroe, Ivy Research Council

Acknowledgements

Firstly, I would like a say a very special thank you to my wonderful partner Brylea Hollinshead with whom daily impassioned discussions of this thesis instilled within me a passionate, dazzling, and wondrous sense of excitement for this topic that will last me a lifetime.

To my primary supervisor Andrew Vonasch. Though your perfectionist qualities made for endless cycles of revisions that were often the cause of great mental turmoil I am a far better writer and researcher because of it. Thank you for pushing me. I also thank you, and my secondary supervisor Andrew Monroe, not only for the outstanding quality of supervision you provided me with but also for the friendly manner in which it was provided. To the moral psych lab group (particularly Scott Danielson and Natasha Dore for their friendship in and outside of the lab) thank you for all the animated discussions that have contributed to this thesis in incalculable ways.

Throughout my time at the University of Canterbury, the teacher's assistants and lecturers I had instilled within me a sense of passion and wonder for the world and ways to think about and investigate it. I am particularly thankful to Roseanna Brailsford and Michael-John Turp for their especially passionate, engaging, and thought-provoking styles of teaching. You are my teaching role models and I hope I can be for other students what you were to me.

Lastly, I would also like to express my extreme gratitude to all my family and friends. To George, my best mate and pet, I thank you for the years of friendship and comfort you provided me. I miss you dearly. Finally, to my parents and grandparents, thank you for raising me. The grit, courage, compassion, workaholism, and can-do attitude you have modelled and instilled within me have provided me with the necessary traits to pursue projects in and outside of academia that fill my life with joy and meaning.

Forever Thankful!

Table of contents

List of Figures

List of Tables

Abstract

Judgements of whether a person intentionally caused harm are consequential—it can mean the difference between murder and manslaughter, for example. But how do people determine whether someone caused harm intentionally? According to The Trade-off Justification Model (TJM) people who cause harm without justifiable reasons are unintentionally signaling that harm was intentional because people without harmful intentions would not willingly decide to cause harm without sufficient reason. Consequently, people judge unjustified harms as intentional. Eight preregistered experiments (N = 1621) tested and extended this claim. Across a wide range of scenarios (Experiment 1) and cultures (Experiments 2a-3b) people tended to judge unjustified harms as intentional but justified harms as not intentional. Moreover, people changed their mind about whether a harm was intentional after learning reasons that made the harm justified or unjustified (Experiment 4). The last two experiments extend this claim by investigating how people judge whether a harm was justified. We show that people's cultural (Experiment 5) and group (Experiment 6) identities set the norms/values they use to evaluate the justifiability of causing harm, thereby leading to different judgements about whether the harm was intentional. People's perceptions of others are thus informed by their personal identities (which group/cultural membership is a part of). We close by discussing the implications of this for science—what morality's theoretical role in this process is—and society—how individual differences in values and culture may create misunderstandings of whether a harm was intentional and cause conflicts between different political groups.

Keywords: Attributions, Intentionality, Trade-off Justification Model, Signaling Theory, Individual Differences

A Sign of ill Intent: Unjustified Harming Signals Harm was Intentional

On June 4, 2020 in Buffalo, New York, police were called in to control protests about George Floyd's death. While clearing an area of protesters, a police officer shoved 75-year-old Martin Gugino to the ground. He suffered a head injury when he hit the ground, was unable to walk for two weeks, and was hospitalized for 27 days (Rose & Levenson, 2020, June 16; Rose et al., 2020, June 30). There is no doubt the police officer caused harm, but people disagree whether it was intentional. Some people think the police officer did not intentionally harm Mr. Gugino—they intended to clear the area and accidentally used too much force on the old man. Other people think the harm was intentional—the police officer shoved Mr. Gugino too hard without any justifiable reason for using such aggressive force. This difference in perceived intentionality is important: If it was intentional, the shove was a clear crime: assault and battery; but if it was an accidental side-effect of following departmental procedures, there was no crime. Moreover, public reaction also depends on the intent—people who think it was intentional want the officer to be punished.

Mr. Gugino's example highlights the importance of understanding how people determine whether someone caused harm intentionally. With the importance of these judgements in mind, the topic of the current research is, "how do people determine whether someone caused harm intentionally?" Specifically, this research focuses on testing a pathway specified by the Trade-off Justification Model for how people determine whether a harm was caused intentionally. The Trade-off Justification Model (TJM) argues that people look at the justifiability of the harm to determine whether it was intentional. In essence, people consider whether there were any reasons (apart from intending the harm itself) sufficient to explain the person's willingness to cause harm. If no other sufficient reason is found, then it shows that the person was willing to cause the harm, had no other justifiable reason for causing it, yet decided to cause it anyway. So, people think the harm was what the person intended and,

consequently, attribute it as intentional. Thus, in Martin Gugino's case, TJM predicts that those who see the police officer's actions as a justified use of force in dispersing potentially violent protestors will see the harm as not intentional, but people who see the officer's actions as unjustified brutality in the service of putting down legitimate protests will perceive the same actions as clearly intentional.

The predictions of TJM go far beyond Mr. Gugino's situation: We conducted eight preregistered experiments (including two replication studies) testing TJM across situations ranging from CEOs causing environmental degradation, to parents causing their child misery through extra schooling, to building extensions blocking the neighbors' views. Moreover, we believe this mechanism should apply across multiple cultures, therefore, we tested it in six countries: US, UK, China, Poland, South Africa, and New Zealand. We even explored how internalizing the norms and values of the groups and cultures people identify with can affect the perceived justifiability of causing a harm and how this can lead to differences in the perceived intentionally of a harm. Overall, we found support for the core prediction of TJM: that judgements of whether a harm was intentional are tied to the perceived justifiability of causing the harm. We conclude that intentionality judgements depend on whether the perceiver judges the harm to be justified and discuss the relevance of this insight for science and society.

Attributing Intentionality

People attempt to determine whether another's actions were the product of personal or impersonal causes, driven by the self or the situation, and, crucially, whether they were intentional or accidental (Gilbert, 1998; Heider, 1985; Jones & Davis, 1965; Jones & Nisbett, 1972; Malle, 2006; McArthur, 1972). Determining whether an action was intentional, however, is inherently challenging because people do not have direct access to others' mental states (Pronin et al., 2004)—the minds of others are invisible. Yet, children as young as 12-18

months seem to have some insight into the minds of others (Liszkowski et al., 2006; Moll & Tomasello, 2007), by 14-18 months distinguish between accidental and intentional action (Carpenter et al., 1998), by 25-months-old anticipate others' intentions (Southgate et al., 2007), and by 3-5 years old incorporate intentions into their moral judgements (Li et al., 2020; Li & Tomasello, 2018). Given that the minds of others are invisible, what are people's judgements based off? Here, we first review two of the predominant answers to this question before positing our own.

Viewing people as naïve scientists, the "theory-theory" theory of mind was one of the earliest dominant theories to provide insight into how people determine whether something was intentional (Gopnik & Wellman, 1992). The theory-theory argues that people reason about the causes of others' actions like naïve scientists—positing or recruiting abstract constructs (mental states) to understand the unobserved underlying causes of behavior and to predict how people will act in the future (Spaulding, 2018). For example, John may witness an officer cause a protestor harm and infer whether it was intentional by determining which possible combination of mental states (beliefs, desires, etc.) are most plausible for the observed behavior and whether the posited mental states satisfy the conditions of intentional action (Malle & Knobe, 1997).

Recently, however, this view of person as naïve scientist has come under attack. Knobe (2003, 2010) has argued that people do not reason about intentions like rational, amoral scientists. Rather, moral considerations figure into whether people consider something intentional. Consequently, the relationship between intentionality judgements and moral judgements is bi-directional—intentionality judgements can influence judgements of praise and blame, and judgements of praise and blame can influence intentionality judgements. Consider the following case as an example:

> The vice-president of a company went to the chairman of the board [(CEO)] and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed" (Knobe, 2003, p.191).

Though it could plausibly be viewed as an unintentional side-effect of trying to make profit, when people were presented with this scenario 82% said that the chairman/CEO intentionally harmed the environment. When the word harm is replaced with help, however, only 23% of people said the chairman intentionally helped the environment. This phenomenon in which people typically judge harmful side-effects as more intentional than similarly helpful side-effects is known as the side-effect effect (Sauer, 2014) and is often used in support of the view that morality biases or influences intentionality judgements (Cushman & Mele, 2008; Nadelhoffer, 2004, 2006; Pettit and Knobe, 2009; Sauer & Bates, 2013). On this view, people judge whether something was caused intentionally based partially on the moral goodness/badness of the outcome.

The view that morality biases or influences intentionality has also come under great criticism, both on theoretical and methodological grounds (Cova et al., 2016; Guglielmo & Malle, 2010; Hindriks, 2014; Holton, 2010; Laurent et al., 2015; Machery, 2008; Uttich & Lombrozo, 2010). We join the critics in arguing that it is not morality per-say that is influencing people's judgements of intentionality. Rather, we argue that people are using social information, particularly information about the perceived justifiability of the harm, to infer whether it was intentional. In the following section we outline our reasons for this claim by drawing on insights from signaling theory.

Signaling Theory, Social Information, and Justifiability

Signaling theory is oriented around how certain behaviors of organisms can reduce information asymmetry (Connelly et al., 2011; Spence, 1973). That is, if I know something you do not, how my behaviors can show what I know. There are thus three main parts of signaling theory—the signaler, the signal, and the receiver. Most research on signaling theory focuses on how signalers intentionally communicate (signal) positive, imperceptible qualities of themselves to others (receivers) (e.g., Ahlers et al., 2015; Karasek & Bryant, 2012; Moss et al. 2015). By comparison, in the cases we are interested in, a person signaling that they have purposefully caused harm is not a positive quality for them to communicate. Moreover, we also think that the way a person is signaling they have intentionally caused harm is not something they are necessarily doing on purpose. Thus, we are not using signaling theory to understand how a person might intentionally signal imperceptible positive aspects of themselves, but rather how a receiver might detect signals of others' negative unobservable aspects through unintended side-effects of their actions. Understood in this way, we argue that inferring whether a person intentionally caused harm can be conceptualized within a signaling framework: the minds of others are invisible, so in order to discern whether someone intentionally caused harm people look at the signals provided by the agent's actions within the social context.

There are many challenges associated with relying on signals, however, as they can vary in their reliability; while some may provide accurate information about the individual's unobservable features (e.g., their mental states) and reduce information asymmetry, others are red herrings—their function is to mislead others (e.g., mass and fighting ability are not reliably correlated, yet some species can bluff their way out of a fight by appearing bigger and reap undue rewards, Backwell et al., 2000). Using signals to determine whether something was intentional is thus no straightforward task—people must be careful what signals they base their judgements off.

Costly signaling (Bird & Smith, 2005; Grafen, 1990; Zahavi, 1975), however, provides a solution to the problem of deception and greatly simplifies this task by indicating what signals people should pay attention to. At its core, costly signaling states that if the costs of signaling the non-directly observable information are high enough that falsification is not worth it, then the signal becomes reliable. The organism detecting the signal can thus have a high credence that the signal is conveying accurate information. A classic example of this is stotting: A gazelle jumping straight high up into the air upon noticing a predator. The costs of stotting are high enough such that weak, unhealthy gazelle cannot do it. Thus, stotting provides a reliable signal to the predator that the gazelle is healthy and not worth chasing. This classic example, however, like most research on signaling, is focused on how something may intentionally communicate a positive, imperceptible quality of itself. By contrast, an example of costly signaling in the way we apply it is Jerrod Murray: Jerrod Murray confessed to murdering fellow student Generro Sanchez on the 6th of December 2017 simply because he wanted to know how it would feel (Knittle, 2017, December 17). The costs of murdering someone simply to know how it feels are high enough that people without callous-unemotional traits do not do it. Thus, murder for the sake of wanting to know how it feels provides a reliable signal that the person is callous and unemotional. We therefore extend the idea that inferences of imperceptible information based on costs are more reliable to negative judgements of others' intentions: when people are trying to infer whether others intentionally caused harm they reasonably give extra weight to signals associated with the person incurring a cost.

Psychologists have already applied signaling theory to a range of mental state and trait inferences. Most notably, inferences of altruistic motives and traits (Hardy & Van Vugt, 2006; Ogunfowora et al., 2018; Van Vugt & Hardy, 2010), sincerity (Ohtsubo & Watanabe, 2009), and whether the person has traits of a high quality (Spence, 1973; Murphy, 2019; Park

et al., 2019). What we propose that is novel, is that many key findings in the attribution literature can be conceived of as people basing their judgements off social information that signals whether others caused harm intentionally.

The concept of social information is broad, it refers to the reduction of uncertainty that can occur through observing an organism's behavior in a social context. That is, how knowing the social norms etc. that a person is acting within can provide others with insight that reduces uncertainty about the underlying motives etc. they may have behind their actions (Detweiler, 1975). Social information can thus come in a variety of forms (Guglielmo, 2015) and provide numerous different signals that observers can detect. For example, a person's willingness to violate social or moral norms can signal counter-normative mental states (Holton, 2010; Uttich & Lombrozo, 2010); simply causing harm can signal the person knew harm would occur (Beebe & Buckwalter, 2010; Beebe & Jensen, 2012; Nakamura, 2018); a person who does not care about a morally relevant effect of their actions can signal they are especially blameworthy (Guglielmo & Malle, 2010; Hindriks et al., 2016); a person's bad character can signal they are likely to cause harm (Phelan & Sarkissian 2008; Pizarro & Tannenbaum 2012; Uhlmann et al., 2015); and a willingness to make certain trade-off can signal intentions (Machery, 2008).

The Trade-off Justification Model (TJM) (Rowe et al., 2020; Vonasch & Baumeister, 2017) argues that the unjustifiability of causing the harm is a particularly important piece of social information—unjustified harming is a clear signal that the harm was intentional.

Arguably all actions cause some harmful side-effects—police speeding to a crime scene emit harmful emissions, for example. However, according to TJM people would think that *that* side-effect was unintentional because the benefits of driving to the crime scene greatly outweigh the minor environmental costs—the harm was so minor that the police officer probably did not even consider it when deciding to race off to the crime scene.

Contrast this with situations where the costs greatly exceed the benefits. Costly, severe, and salient harms are such important pieces of social information that it is unlikely a person did not consider it in their decision making process—for example, an ambulance worker deciding which of two lives to save after an accident, one of whom is their child, cannot avoid considering that treating the stranger first may result in the death of their child. That someone considered the harms and knew they were making such a trade-off, however, is not sufficient to infer intentionally, for, as Thomas Aquinas famously argued, people can knowingly cause others' deaths yet still not be judged as intending it (Aquinas, 13th c, II-II, Qu. 64, Art.7; Phelan & Sarkissian, 2009). This is where justifiability is key: If the reasons for causing the harm do not justify the costs, then it signals that the person considered/knew that harm would occur, had no justifiable reason for causing it, yet decided to cause it anyway. Consequently, signaling that the harm was intentional—the costs of causing the harm are high enough (i.e., costs > benefits) that people without harmful intentions would not do it. In contrast, when the harm was caused justifiably there is a lot more noise associated with the signal: if the reason for causing the harm outweighs the costs then the person still likely knew that harm would occur, but uncertainty remains about whether the person intended the harm or did it because of the benefits. Thus, in stark contrast to Machery's trade-off model (2008), the Trade-off Justification Model reasons that not all costs will be judged intentional—only unjustified ones.

In summary, how people determine whether someone caused harm intentionally can be conceptualized within a signaling framework: people use the signal provided by the person causing harm without justifying benefit to determine whether it was intentional—if the reasons for causing the harm do not justify the costs then it signals that the harm was intentional. The core claim of TJM is therefore that unjustified harms will be judged intentional but not justified ones.

Evidence Supporting TJM

The previous section outlined the theoretical basis of TJM. This section reviews existing evidence supporting it.

Both direct and indirect tests of TJM have lent support for the model's core claim. Two studies have directly tested the model's claim; Vonasch and Baumeister (2017) tested the claim that taboo trade-offs (i.e., trade-offs between sacred and non-sacred goods, or between things that are assigned infinite value and things that are not—e.g., love and money; Tetlock, 2003; Tetlock et al., 2000) would be judged as highly intentional due being highly unjustified. They found strong support for their claim with 95% of people judging that the harmful side-effect resulting from the taboo trade-off was intentional (Experiment 1). Rowe et al., (2020) tested the claim that the role a person occupies can increase or decrease perceptions that harm was intentional due to how it decreases or increases the justifiability of the role occupant's actions. They also found strong support for their claim: when occupying a role made the harm less justifiable it was judged more intentional (Experiments 1 and 2). Conversely, when occupying a role made the harm more justifiable it was judged less intentional (Experiment 3). While the number of studies directly testing TJM are few, their results are clear: justifiability matters.

Although only two studies have intentionally manipulated the justifiability of the harm, there are many cases which have unintentionally altered it as a side-effect of changing another variable (e.g., Mele & Cushman's pond case, 2007; Phelan & Sarkissian's city planner case, 2008). One of the best pieces of evidence for TJM comes from one such incidental manipulation by Phelan and Sarkissian (2009). Directly supporting TJM, their results showed that the relationship between the stated main goal and the harmful side-effect greatly affected judgements of whether the harmful side-effect was perceived as intentional: when the main goal (taking a hill for fun in a battle) did not justify the harmful side-effect

(soldiers dying), people typically judged the harm as intentional (73.5% of people judged it as intentional when averaging across variations 2&4). However, when the main goal (taking an important hill in a battle) meant the lieutenant was more justified in causing the harmful side-effect, people were far less likely to judge it as intentional (47.5% of people judged it as intentional when averaging across variations 1&3). Contrast this large change (a 26% difference across conditions) in attributions caused by increasing the justifiability of the harm, with the minor change in attributions caused by varying the lieutenant's words: whether the lieutenant said they did or did not care about their soldiers' lives hardly caused any change (only a 5% difference across conditions).

In summary, TJM's core claim that unjustified harms, but not justified harms, are judged intentional is supported by both direct and indirect tests.

Extending the Model

The current state of the evidence supporting TJM and its answer for the question motivating this research (how do people determine whether someone caused harm intentionally?) is dissatisfactory for two notable reasons.

Firstly, while the studies reviewed indicate the validity of the model, they are also limited—the populations sampled WEIRD (Henrich, et al., 2010), the designs restricted (all studies used only between-subjects designs), and the scenarios utilized constrained (e.g., in Rowe et al., 2020 the scenarios involved harms comparatively minor to the typical ones researchers use, and none of the previous studies looked at whether the mechanism for attributing harmful intentions to groups is the same).

Secondly, the model states that people determine whether harm was intentional by looking at whether the harm was justified, but how are people determining whether the harm is justified? There are many possible ways, all consistent with TJM. For instance, the observer may judge whether the harm was justified from the agent's perspective (i.e., the

observer uses the agent's beliefs to judge whether the harm was justified). The observer may also judge whether the harm was justified within the social context (i.e., using the provided social information to judge whether the harm was justified from the perspective of the society in which the actor is acting, or, to borrow Hume's phrase, "the common point of view"; 1740, T 3.3.1.30). Judgements of justifiability could also be made from the observer's own perspective (i.e., the observer uses their own beliefs to judge whether the harm was justified to them). Thus, in the paradigmatic CEO case, a person could (1) judge whether the CEO was justified in causing harm from the CEO's perspective—was the harm justified in the CEO's mind? (2) from the common point of view—was the harm justified within the immediate social context? or (3) from their own perspective—was the harm justified to the observer? Based on which perspective is taken, different judgements of whether the harm is justified, and consequently judged intentional, could be made.

Supporting the view that observers are judging the justifiability of harms from their own perspective are the growing number of experiments in the literature documenting the existence of observer effects—or the effect of the judge's beliefs, traits, and background on attributions of intentions (e.g., Cokely & Feltz; 2009; Liao et al., 2018; Robbins et al., 2017; Voiklis & Nickerson, 2012). For example, Tannenbaum et al., (2007) showed that people's moral values track judgements of intentionality; people who valued the environment more were more likely to say that the CEO intentionally harmed it (80% of people who valued protecting the environment judged the harm as intentional, whereas only 63% of people who did not value protecting the environment judged it as intentional). Investigating how certain aspects of observers may affect their perceptions of the justifiability of a harm (and consequently whether they judge it intentional) has the potential to deepen our understanding of the most interesting aspects of Mr. Gugino's example—why people observed the exact same situation yet disagreed whether the harm was intentional.

Current Research

The previous section highlighted two points of dissatisfaction for the answer provided to the question "how do people determine whether someone caused harm intentionally?" The current research is thus two-fold. Firstly, we further and more rigorously investigated the link between justifiability and intentionality by using a variety of scenarios, populations, and complementary experimental designs that address the limitations previously highlighted (Experiments 1-4). Secondly, we investigated how the dispositional beliefs of observers may affect their perceptions of the justifiability of a harm, leading some people to judge a harm as more intentional than others despite everyone being given the same information (Experiments 5-6).

More specifically, Experiment 1 tests the generalizability of the model by giving people a wide range of scenarios and assessing the relationship between justifiability and intentionality across the 19 scenarios. Experiments 2a-3b compliment Experiment 1 in two ways. Firstly, while Experiment 1 tests the generalizability of the model across situations, Experiments 2a-3b tests the cultural generalizability of the model by sampling from 4 different countries. Secondly, while Experiment 1 is correlational in design, Experiments 2a-3b experimentally manipulate the justifiability of causing harm by varying either the amount of costs or the amount of benefits in the trade-off. Experiment 4 compliments the previous experiments by testing whether people incorporate new justifying information and change their judgements about whether a person intentionally caused harm. Finally, in Experiments 5 and 6 we investigated whether the cultural or group norms and values an observer has internalized affect the perceived justifiability of a harm and, consequently, the perceived intentionality of the harm.

In an effort to promote open, robust science and avoid abusing researchers' degrees of freedom, all experiments were preregistered, include a discussion of power/sample size

where relevant, and have the anonymized data posted online (https://osf.io/u9sp6/?view_only=9cb9b9749dae4a34a72e27ae083c5d50). The cultural generalizability of psychological studies has also been limited for years by being tested on WEIRD populations (Henrich, et al., 2010). With the rise of popular crowdsourcing platforms such as Mechanical Turk and Prolific allowing easy collection of data from people across the world, such a limitation is no longer justified. Here we have taken care to test the cultural generalizability of our findings by sampling people from 6 different countries (UK, US, SA, NZ, Poland, and China). Overall, the results support TJM's core hypothesis that people's judgements of whether others intentionally caused harm tracks the perceived justifiability of the harm and that this is true across a variety of scenarios and cultures. Results also support extensions of the model based on how aspects of the observer can affect the perceived justifiability, and, consequently, the intentionality of the harm.

The materials for all experiments were approved by the Human Research Ethics Committee at the University of Canterbury (HEC#2019/02).

Experiment 1

Causing harm without justified reason has been claimed to signal that the harm was intentional. Experiment 1 aimed to test the generalizability of this hypothesized relationship between justifiability and intentionality across a range of scenarios. Most of the vignettes in the literature have been restricted to a common situation about a CEO (or person of equivalent status) who implements some sort of program that has a harmful side-effect. By using a wide range of scenarios—scenarios where the focus of the attribution is about a political group or movement, for example—we can test whether the hypothesized mechanism applies in numerous contexts.

Method

Preregistration https://aspredicted.org/blind.php?x=zc4jx6Procedure

Procedure

Participants read 19 different vignettes, presented in random order, about agents who caused harm. After reading each vignette, participants answered questions about whether they thought the harm was intentional and whether they thought the harm was justified. The harm, the agent who caused the harm, as well as the justifiability of causing harm are varied throughout the vignettes (see example below, all vignettes are available in Appendix A). We took care to ensure that the harms were those people would typically agree were unjustified, somewhat justified, or completely unjustified as we are interested in testing the effects at the vignette level first before moving on to test the effects of justifiability at the individual level later.

Participants

40 undergraduate psychology students from Appalachian State University were requested to participate in this study for course credit. Due to a technological error, the system sent through 49 participants. Eight were excluded for failing an attention check, leaving a final sample of 41 participants (M_{age} = 19.0, SD = 1.98). The majority were female (33 Female, 7 Male, 1 Gender diverse) and white (87.8% White, 7.3% Black).

Measures

Perceived intentions and justifiability were measured on a 7-point Likert scale where 1 = Definitely Not, 4 = Maybe, and 7 = Definitely Yes.

Example Measure of Perceived Justifiability. “Were the supporters justified in ending slavery?”

Example Measure of Intentions to Harm. "Did the supporters intentionally harm the economy?"

Attention Check. In online studies (particularly those using popular crowd sourcing platforms), bots and inattentive participants can be a major issue. To avoid this issue, we included an open-ended question as an attention check ("In one to two sentences, please explain why you thought the CEO's decision was/was not justified.") in all our studies to screen out bots and inattentive participants (Chmielewski & Kucker, 2020). Participants were excluded if their open-ended responses did not answer the question—e.g., "was the business there first?" was excluded.

Example Vignette

> The abolitionist movement in Laputa was aimed at ending slavery in the country. However, supporters of this movement also knew that ending slavery would have negative effects on the economy. The movement was successful and, sure enough, there were negative effects on the country's economy.

Results

Analytic Plan

We had observations for 19 different scenarios and were interested in investigating whether the effect of justifying social information was similar across scenarios (rather than individuals). Thus, we used linear mixed models in Jamovi (clustering by vignette and including random slopes and intercepts for each vignette) to estimate the effect of justifying social information on intentions at the vignette level.

Key Results

Perceived Justifiability Predicts Intention judgements. As predicted, perceived unjustifiability predicts attributions of harmful intentions, $F(1, 18.3) = 17.7$, $p < .001$: Harms that were judged as more unjustified were judged as more intentionally caused ($B = -0.23$, 95%CI:[-.33, -.12], $t(18.3) = -4.20$, $p < .001$, $R^2_{conditional} = .22$). The relationship between perceived justifiability and attributions of harmful intentions showed slight variation in intercepts ($SD = 0.69$, $Var = 0.48$) and slopes ($SD = 0.15$, $Var = 0.02$) across vignettes, indicating that the effect of perceived justifiability on intention judgements was slightly different across scenarios (i.e., in some scenarios the effect of justifiability on intention judgements was slightly stronger).

Visual analysis of mean justifiability and intentionality responses also support the hypothesized negative association by showing a clear trend that in vignettes where the harm is judged to be less justified, the harm is perceived as more intentional (Figure 1, Table 1).

Figure 1

Line Graph showing the Negative Association between Mean Intentionality and Mean Justifiability Responses across the Series of Vignettes in Experiment 1, Organized from Low to High Justifiability

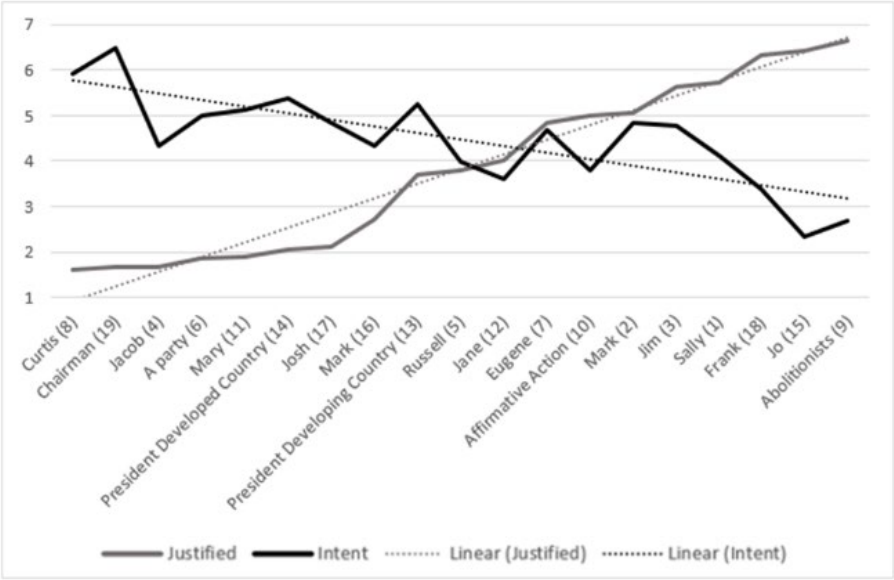


Table 1

Means (and standard deviations) in Experiment 1, Split by Vignette and Organized from Low to High Justifiability

Vignette	Justifiability	Intentionality
Curtis (8)	1.61(1.18)	5.93(1.35)
Chairman (19)	1.66(0.94)	6.49(0.84)
Jacob (4)	1.68(1.29)	4.34(1.65)
A Party (6)	1.85(1.24)	4.98(1.70)
Mary (11)	1.88(1.10)	5.12(1.68)
President Developed Country (14)	2.05(1.34)	5.39(1.41)
Josh (17)	2.10(1.50)	4.83(2.06)
Mark (16)	2.71(1.57)	4.32(1.90)
President Developing Country (13)	3.71(1.81)	5.24(1.41)
Russell (5)	3.78(1.56)	3.98(1.57)
Jane (12)	4.00(1.64)	3.59(1.84)
Eugene (7)	4.85(1.62)	4.68(1.72)
Affirmative Action (10)	5.00(1.87)	3.80(2.03)
Mark (2)	5.05(1.43)	4.85(1.93)
Jim (3)	5.63(1.44)	4.76(2.34)
Sally (1)	5.71(1.12)	4.10(2.27)
Frank (18)	6.32(1.21)	3.39(2.04)
Jo (15)	6.41(1.05)	2.32(1.51)
Abolitionists (9)	6.63(1.04)	2.68(1.60)

Note. The order of the vignettes was randomized. The numbers next to the name of each vignette are simply to aid in identifying each vignette (available in the appendix).

Exploratory Analyses

We were also interested in testing individual differences, i.e., whether the effect of perceived justifiability on intention judgements is consistent across participants. To test this we conducted another linear mixed model in Jamovi (clustering by participant and including random slopes and intercepts for each participant) to estimate the effect of justifying social information on intentions at the participant level. As expected, perceived justifiability predicted attributions of harmful intentions at the participant level, $F(1, 40.4) = 114$, $p < .001$: The harms that people judged as more unjustified were judged as more intentionally caused

(B = -0.36, 95%CI:[-.42, -.29], t(40.4) = - 10.7, p < .001, $R^2_{conditional}$ = .39). The relationship between perceived justifiability and attributions of harmful intentions showed some variation in intercepts (SD = 0.95, Var = 0.90) and some variation in slopes (SD = 0.13, Var = 0.02) across participants. We interpret the variation in intercepts and slopes as meaning that some people are less responsive to justifying social information than others and that some people require the harm to be more justified before they consider it as not intentional.

Discussion

Experiment 1 showed that the general relationship between justifiability and intentionality was as predicted: people tended to judge less justified harms as more intentional (Figure 1). Moreover, this trend emerged at both the vignette level and individual level: harms that people tend to see as unjustified are the ones that people tend to judge intentional. Also, individuals who tended to see harm as less justified in general tended to judge harm as more intentional.

Experiment 1 supports the generalizability of TJM across various contexts, addressing one of the limitations of previous tests of the model. However, no tightly controlled manipulations took place—while the scenarios were designed to vary in their justifiability, there were also other things that varied between them (such as whether the attribution was made at a group or an individual). While we suspect the other variations did not drive the trend, more controlled experiments need to be done to be certain. Thus, Experiments 2a-3b set out to isolate the manipulation of one variable (justifiability) and test (and replicate) the effects of this manipulation across cultures.

Experiment 2a

Experiment 1 verified that justifiability and intentionality were linked across a wide range of scenarios. Experiment 2a aimed to provide a more controlled test by manipulating

only one variable—the justifiability of the CEO running their program. We manipulated the justifiability of causing the harm by varying how much profit a CEO made as a result of running a program. However, we could not use the harm in the original CEO vignette (environmental destruction) because it is taboo and no increase in profit can justify a taboo trade-off (Tetlock, 2003; Tetlock et al., 2000; Vonasch & Baumeister, 2017). Thus, we changed the harm to something comparatively minor—the CEO annoying their neighbors by expanding their building and blocking their view. We predicted in scenarios where the program made more profit people would judge the decision to run the program as more justified and perceive the harmful side-effect (annoying the neighbors) as less intentional. Moreover, we predicted that the relationship between the amount of profit and perceived harmful intentions would be mediated by the perceived justifiability of running the program. Specifically, we predicted that when the CEO's company made more profit people would judge the decision as more justified, and that people who thought the decision was more justified would judge the harm as less intentional.

This Experiment also included an exploratory component: we wanted to begin investigating how observers are judging the justifiability of the harm. Based on previous research demonstrating observer effects (the effect of the judge's beliefs, traits, and background on attributions of intentions; see, for example Tannenbaum et al., 2007), we hypothesized that observers' perceptions of whether the act was morally wrong impacts intentionality judgements via its effect on the perceived justifiability of the harm. We predicted that, as immoral actions are by their nature hard to justify, observers who view the decision as morally wrong will be more likely to judge it as unjustified, and, therefore, as more intentional.

Method

Preregistration https://aspredicted.org/blind.php?x=4gk9qa

Procedure

Participants read one of three randomly assigned vignettes (see below) and then answered questions about the justifiability of the CEO's decision, the CEO's intentions, the moral wrongness of the CEO's actions, and an attention check. The order of the justifiability and intention questions were randomized to avoid possible order effects (none were found). Participants then completed demographics (gender, age, race) and were debriefed.

Participants

As preregistered, we requested 300 UK participants from Prolific per power analyses indicating 250 participants are required for sufficient power (95%) to detect a medium effect size in a one-way ANOVA (the average effect size in social psychology). We oversampled to account for potential exclusions/drop-out. Prolific sent 317 participants. 18 participants were excluded for failing an attention check, leaving a final sample of 299 participants (194 Female, 102 Male, 3 declined to state) (M_{age} = 36.9, SD = 11.5).

Measures

The measures of intentionality and justifiability were the same as in the previous experiment, but with slight changes in wording to fit the vignette (e.g., "Was the CEO's decision justified?", **"**Did the CEO annoy the neighbours on purpose?").

This experiment also included two new measures—moral wrongness (measured via a dichotomous yes/no response) and intentionality attributions for concrete behaviors (measured on the same 7-point Likert scale as perceived justifiability and harmful intentions).

Moral Wrongness. "Were the CEO's actions morally wrong?"

Intentionality Attributions for 'Concrete' Behaviors. Whereas attributions of abstract 'high level' behaviors (e.g., intentionally causing harm) have been shown to be affected by social information, attributions of concrete 'low-level' behaviors (e.g., picking up a pen) have been shown to be unaffected by social information (Monroe et al., 2015). We thus included a measure of intentionality for concrete behaviors (**"**Did the CEO run the program on purpose?") to highlight the difference between the types of attributions that are affected by social information and those that are not.

Vignette

> "The vice-president of a company went to the CEO and said, "We are thinking of hiring more staff and expanding our office building by increasing it 3 stories. It will help us increase profits by ***fifty pounds/fifty thousand pounds/five million pounds*** per year, but it will also really annoy the neighbors by blocking their view."
>
> The CEO decided to run the program. Sure enough, the neighbors view was blocked and they were really annoyed."

Results

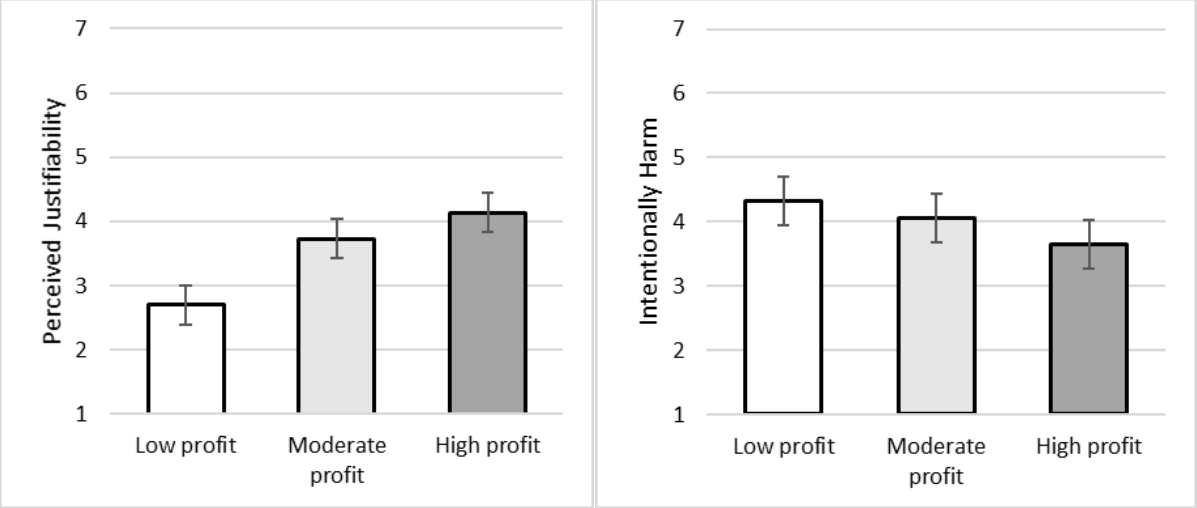
Key Results

Perceived Justifiability of the Decision. As predicted, a one-way ANOVA revealed a significant main effect of profit on the perceived justifiability of the CEO's decision, $F(2,296) = 23.3$, $p < .001$, $\eta^2 = .136$). Bonferroni corrected post-hoc comparisons showed that the change in means was as we predicted: when the program's profit was low ($M = 2.70$, $SD = 1.56$) people judged the CEO's decision as less justified than when the profit was moderate ($M = 3.73$, $SD = 1.34$), $t(296) = 4.78$, $p_{bonferroni} < .001$, $d = 0.27$, or when the profit was high ($M = 4.14$, $SD = 1.65$), $t(296) = 6.64$, $p_{bonferroni} < .001$, $d = 0.38$ (Figure 2). Though not

significantly different ($p_{bonferroni}$ = .162) the pattern of means for moderate compared to high are in the predicted direction (High < Moderate) (Table 2).

Figure 2

Mean Perceived Justifiability and Intentionality Responses in Experiment 2a by Condition



Note. Perceived intentions and justifiability were measured on a 7-point Likert scale where 1 = Definitely Not, 4 = Maybe, and 7 = Definitely Yes. Error Bars = 95% CI

Table 2

Means (and standard deviations) in Experiment 2a by Condition

	Low profit	Moderate profit	High profit
Perceived Justifiability	2.70(1.56)[a]	3.73(1.34)[b]	4.14(1.65)[b]
Intentionally Harm	4.33(2.05)[a]	4.05(1.93)[ab]	3.65(1.77)[b]
Intentionally run programme	5.37(1.83)[a]	5.34(1.88)[a]	5.49(1.70)[a]

Note. In rows, means with different superscripts are significantly different (p < .05). Differences in means were tested using Bonferroni corrected planned comparisons.

Perceived Intentions to Harm. As predicted, a one-way ANOVA revealed a significant main effect of profit on peoples' judgements of the CEO's intentions to annoy the neighbors, $F(2,296) = 3.14$, $p = .045$, $\eta^2 = .021$). Bonferroni corrected post-hoc comparisons revealed that the change in means was as predicted: When the program's profit was low people perceived the harm as more intentional than when the program's profit was high, $t(296) = -2.49$, $p_{bonferroni} = .040$, $d = 0.14$ (Table 2, Figure 2). Again, whilst not significantly

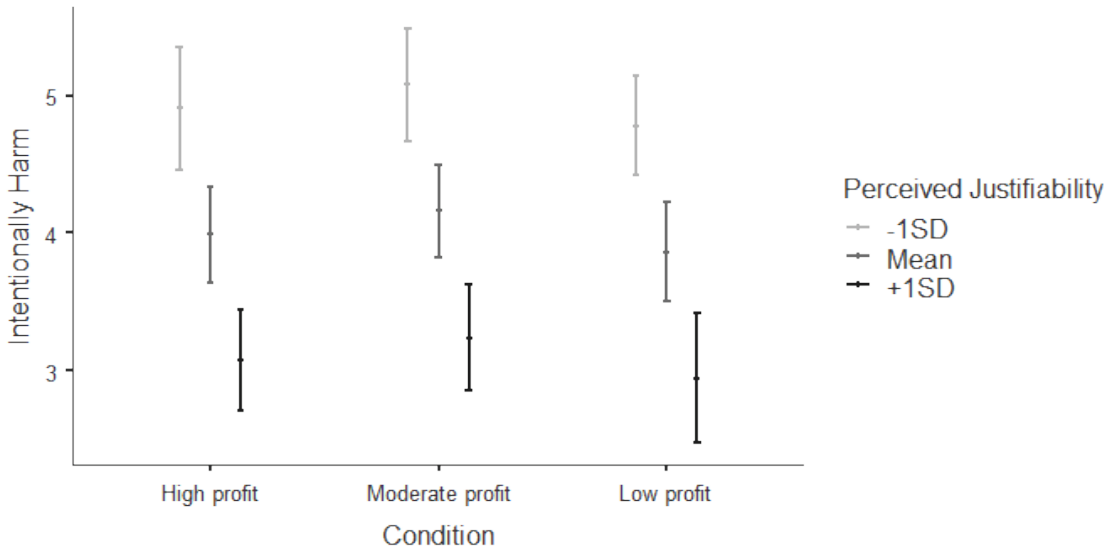
different, the patterns of means for moderate compared to large ($p_{bonferroni}$ = .414) and small compared to a moderate amount ($p_{bonferroni}$ = .905) is in the predicted direction (High < Moderate < Low).

The differences in perceived intention to harm in low vs high profit reflect the perceived differences in the justifiability of the CEO's decision: in the condition where the decision was viewed as most justified, the harm was seen as the least intentional. Incongruently, despite there being a significant difference in the perceived justifiability of the CEO's decision between the low and moderate profit conditions, there was not a significant difference in perceived intentions to harm. However, despite not all being significantly different from one another, the group means are presenting in a linear fashion as predicted (High < Moderate < Low, Table 2). Indeed, exploratory follow up analyses revealed a significant linear trend B = 0.48, 95%CI:[0.10, 0.87], $t(296) = 2.49$, p = .013). Showing that as the profit became lower, people were more likely to judge the harm as intentional.

Furthermore, strongly supporting TJM, the results show a consistent pattern across conditions that participants who thought the decision was unjustified thought the harm was intentional, whereas those who thought the decision was justified thought the harm was not intentional (Figure 3). Indeed, a linear regression showed a strong negative relationship between perceived justifiability and perceived harmful intentions, B = -.55, 95%CI:[-.67, -.43], $t(298) = -8.90$, $p < .001$.

Figure 3

Mean Attributions of Harmful Intentions in Experiment 2a, Split by High/Low Perceived Justifiability of the CEO's Decision



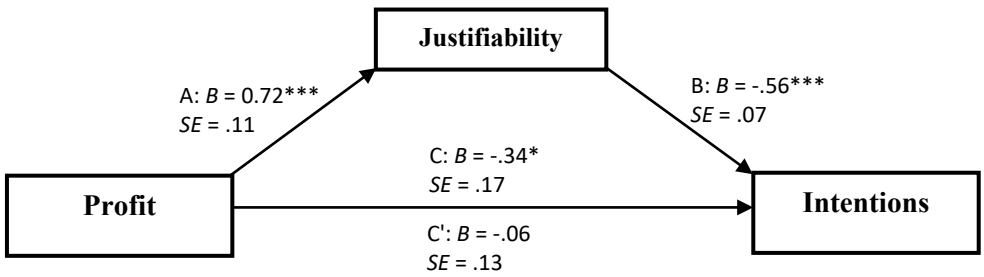
Perceived Intentions to Run the Program. As predicted, a one-way ANOVA showed that how much profit the program made had no effect on perceptions of the CEO's intentions to run the program, $F(2,296) = 0.206$, $p = .814$, $\eta^2 = .001$) (Table 2). A linear regression also supports this by showing that the perceived justifiability of the CEO's decision does not predict perceived intentions to run the program, B = -0.10, 95%CI:[-0.23, 0.02], $t(297) = -1.62$, $p = .106$). Thus, consistent with previous research, judgements of intentions to do concrete actions (e.g., run a program) were much less reactive to social information than judgements of intentions to cause abstract outcomes (e.g., annoy the neighbors).

Pathway from Profit to Perceived Intentions to Harm via Perceived Justifiability. Mediation using the MEDMOD package in JAMOVI and 5000 bootstrap samples revealed that increases in the profit the program made was associated with decreased perceived intentions to harm (annoy the neighbors) and that an increase in the perceived

justifiability of the CEO's decision mediated this link (indirect effect: -0.40, 95%CI:[-0.56, -0.23], SE = 0.08, $p < .001$; Figure 4). Increase in profit was associated with increased perceptions that the CEO's decision was justified ($B = 0.72$, 95%CI:[0.50, 0.94], SE = 0.11, $p < .001$). An increase in the perceived justifiability of the CEO's decision was associated with a decrease in perceptions that the CEO intentionally annoyed the neighbors ($B = -0.56$, 95%CI:[-0.67, -0.42], SE = 0.07, $p < .001$). The remaining direct effect was not significant ($B = -0.06$, 95%CI:[-0.20, 0.32], SE = 0.13, $p = .664$) indicating that the effect was fully mediated. Mediation analyses thus support the hypothesis that it is how the increase in profit affected the perceived justifiability of causing harm that mattered for intention judgements, not something about profit in itself. However, a qualification is warranted here: When mediators are measured at the same time as dependent variables, mediation cannot establish causality (Bullock et al., 2010); nonetheless, results are consistent with the proposed causal model.

Figure 4

Mediation Model Depicting the Pathway from Increased Profit levels to Reduced Perceived Intentions to Harm via Increased Perceived Justifiability in Experiment 2a



Exploratory Analyses

Pathway from Perceived Moral Wrongness to Perceived Harmful Intentions via Perceived Justifiability. We hypothesized that observer's perceptions of whether the act was morally wrong would impact intentionality judgements via its effect on the perceived

justifiability of the harm. An exploratory mediation analysis supported this hypothesis: Observer's judgements that the CEO's actions were morally wrong was associated with increased perceived harmful intentions and perceived justifiability of the CEO's decision mediated this link (indirect effect: 0.82, 95%CI:[.49, 1.14], SE = 0.18, $p < .001$).

Discussion

The results from Experiment 2a strongly support TJM's hypothesis that people's judgements of others' harmful intentions track the social information available: People who thought the harm was *unjustified* judged it as *intentional*, whereas people who thought the harm was *justified* judged it as *not intentional* (Figure 3). Moreover, as the profit became lower people judged the harm as more *unjustified* and more *intentional* (Table 2). Experiment 2a also highlights the limitations social information has on certain types of attributions—whereas social information is highly influential in whether abstract behaviors are judged intentional, it has no effect on whether concrete behaviors are judged intentional.

Experiment 2a also provided preliminary evidence for how observers judge whether the harm was justified: people who judged the harm as morally wrong judged it as more unjustified and, consequently, intentional. Thus, indicating that observers are, at least in part, judging the justifiability of the harm from their own perspective. Admittedly, however, this experiment was not ideally situated to test how observer's judge the justifiability of the harm. Thus, in the following replication studies we dropped the moral wrongness measure and took a different approach in Experiments 5 and 6 by utilizing existing differences in people's cultural backgrounds and political ideologies.

Experiment 2a was more tightly controlled than Experiment 1. Here, only one variable (profit) changed between the vignettes. This subsequently affected people's perceptions of the justifiability of running the program and, consequently, the intentions they

attributed to the CEO. Thus, it provides strong support for TJM's core claim that justifiability affects intentionality judgements. However, it is just one study with one vignette on a WEIRD (Henrich et al., 2010) population. To test the generalizability of the findings we conducted another study using a culturally distinct population (South African compared to British).

Experiment 2b

Experiment 2b aimed to replicate and extend Experiment 2a by using a conceptually identical vignette and distinct population. Participants read a short vignette about a CEO who, through upgrading to an automated assembly line, made 3 workers redundant (retrenched). The CEO made various amounts of profit for upgrading, and, as in Experiment 2a we predicted that as the amount of profit increased that people would judge the CEO's decision as more justified and the harm less intentional. Furthermore, that the relationship between the amount of profit made and perceived harmful intentions would be mediated by perceived justifiability.

Method

Preregistration https://aspredicted.org/blind.php?x=dn5kr4

Procedure

Apart from requesting South African participants, dropping the moral wrongness question, and using a different but conceptually identical vignette the procedure was identical to Experiment 2a.

Participants

We requested 300 South African participants from Prolific, and Prolific sent 313. Twenty seven were excluded for failing an attention check, leaving a final sample of 286

participants (M_{age} = 29.4, SD = 9.53); the majority of participants were female (161 Female, 124 Male, 1 non-binary) and white (47.6% White, 29.0% Black, 8.7% Indian, 14.3% Other).

Measures

The measures were the same as in Experiment 2a but with slight changes in wording to fit the vignette (e.g., "Did the CEO intentionally make the 3 workers unemployed?").

Vignette

> "The vice-president of a company went to the CEO and said, "We are thinking of updating to a new automated assembly line. It will help us increase profits by **500/700,000/2,000,000** Rand per year, but will cause 3 workers to be retrenched." The CEO decided to run the program. Sure enough, 3 of the company's workers were retrenched."[1]

Results

Key Results

Perceived Justifiability of the Decision. As predicted, a one-way ANOVA revealed a significant main effect of profit on the perceived justifiability of the CEO's decision, $F(2,283) = 8.78$, $p < .001$, $\eta^2 = .058$). Also as predicted, a linear contrast showed a significant linear trend, $B = -0.57$, $p < .001$, indicating that as profit decreased, people become less likely to judge the decision as justified. However, a visual analysis of the means and Bonferroni corrected follow up comparisons qualify this result (Figure 5, Table 3). Most notably, the average perceived justifiability in the high profit group (M = 4.80, SD = 1.57) was not greater than the moderate profit group (M = 4.97, SD = 1.47, $t(283) = -0.52$, $p_{\text{bonferroni}} = 1.00$).

[1] To ensure the vignette was written in the culturally appropriate vernacular we consulted with one of the authors South African friends.

Though, as expected, the average perceived justifiability in the moderate profit group was higher compared to the low profit group ($M = 4.05$, $SD = 1.88$, $t(283) = 3.35$, $p_{bonferroni} = .003$, $d = 0.56$).

Table 3

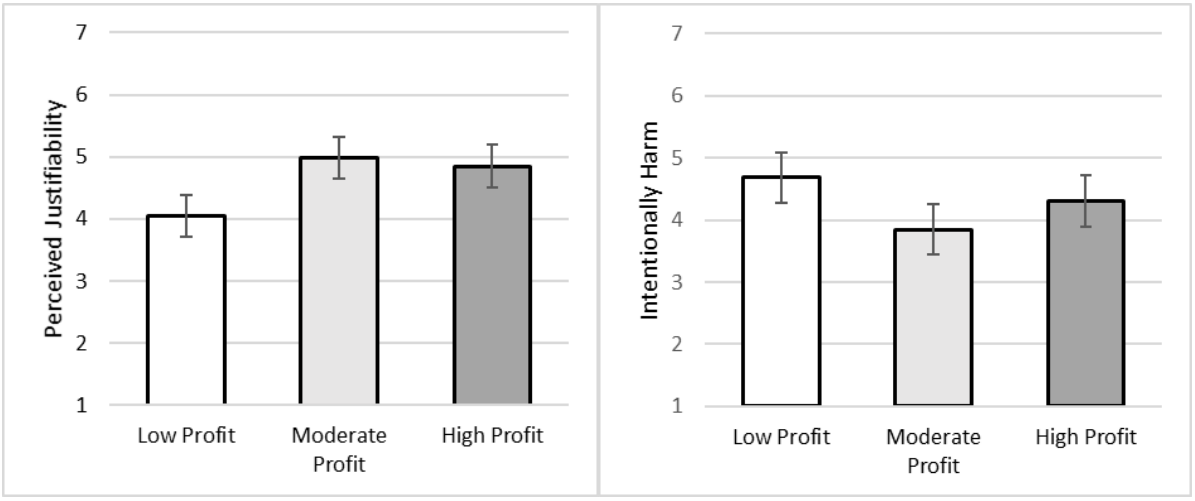
Means (and standard deviations) in Experiment 2b, Split by Condition

	Low Profit	Moderate Profit	High Profit
Perceived Justifiability	4.05(1.89)[a]	4.98(1.49)[b]	4.85(1.55)[b]
Intentionally Harm	4.68(2.07)[a]	3.85(1.98)[b]	4.31(2.15)[ab]
Intentionally Run Programme	5.52(1.91)[a]	5.32(2.04)[a]	5.50(1.91)[a]

Note. In rows, means with different superscripts are significantly different ($p < .05$). Differences in means were tested using Bonferroni corrected planned comparisons.

Figure 5

Mean Perceived Justifiability and Intentionality Responses in Experiment 2b by Condition



Note. Perceived intentions and justifiability were measured on a 7-point Likert scale where 1 = Definitely Not, 4 = Maybe, and 7 = Definitely Yes. Error Bars = 95% CI

Perceived Intentions to Harm. As predicted, a one-way ANOVA revealed a significant main effect of profit on peoples' judgements of the CEO's intentions to make the employees unemployed, $F(2,283) = 3.85$, $p = .022$, $\eta^2 = .026$). However, contrary to predictions, a linear contrast revealed that there was no significant linear trend, $B = 0.26$, $p = .216$, rather, there was a significant quadratic trend, $p = .014$ (Figure 5). The reason for the

quadratic trend is presumably due to the manipulation not working as predicted: we expected that as the amount of profit increased that people would judge the CEO's decision as more justified, but, although people did judge the CEO's decision as more justified when the profit was moderate or high compared to low, people did not judge it as more justified when it was high compared to moderate (see section above). Though the manipulation did not work as expected, exploratory analyses indicate that TJM's key prediction that people who judge a decision to cause harm as *unjustified* will judge the harm as *intentional*, whereas people who judge the decision as *justified* will judge the harm as *not intentional* is still supported: people who judged the CEO's decision as definitely unjustified (i.e., reported justifiability of 1, N = 20) judged the harm as intentional (M = 5.15, SD = 2.43), whereas people who thought the CEO's decision was definitely justified (i.e., reported justifiability of 7, N = 44) judged the harm as not intentional (M = 3.64, SD = 2.26).

Perceived Intentions to Run the Program. As in Experiment 2a, we predicted that judgements of intentions to do concrete actions (e.g., run a program) would be much less reactive to social information than judgements of intentions to cause abstract outcomes (e.g., cause employees to be retrenched). Consistent with Experiment 2a, a one-way ANOVA showed that how much profit the program made had no effect on perceptions of the CEO's intentions to run the program, $F(2,283) = 0.331$, $p = .719$, $\eta^2 = .002$).

Pathway from Profit to Perceived Intentions to Harm via Perceived Justifiability. Mediation using the MEDMOD package in JAMOVI and 5000 bootstrap samples revealed that increases in the profit the program made was associated with decreased perceived intentions to make the 3 employees unemployed and that an increase in the perceived justifiability of the CEO's decision mediated this link (indirect effect: -0.10, 95%CI:[-0.19, -0.02], SE = 0.05, p = .022). Increase in profit was associated with increased perceptions that the CEO's decision was justified (B = 0.40, 95%CI:[0.16, 0.65], SE = 0.12, p

= .001). An increase in the perceived justifiability of the CEO's decision was associated with a decrease in perceptions that the CEO intentionally made the 3 employees unemployed (*B* = -0.26, 95%CI:[-0.41, -0.10], SE = 0.08, *p* = .001). The remaining direct effect was not significant (*B* = -0.08, 95%CI:[-0.38, 0.22], SE = 0.15, *p* = .593), indicating that the effect was fully mediated.

Discussion

The manipulation in Experiment 2b did not work as expected. We anticipated that as profit increased, the justifiability of the CEO's decision would increase, but we did not find this pattern. Rather, an analysis of the open-ended responses to the attention check question ("In one to two sentences, please explain why you thought the CEO's decision was/was not justified.") indicated an unanticipated split in people's reasoning that we expect caused the manipulation to fail: some people argued that as profit increased the decision became less justifiable (e.g., "It is not ethical to retrench someone to increase the companies profits. With the extra money that they made he could have kept on the two employees and still make a profit") – after all, if the company is going to make a lot more profit, what reason does the CEO have for making people redundant as opposed to restructuring their jobs? In contrast, others argued that as profits increased, the CEO's decision became more justifiable (e.g., "In the end it is the CEO's responsibility to make sure the company is profitable. His decision, was in my opinion justified.") – after all, it is the CEO's role to make as much profit as possible, thus, the more profit the CEO's decision makes, the more justified it is.

The results still supported TJM despite the manipulation not working as intended. Experiment 2b showed the same key pattern of results as Experiment 2a but with a different vignette and culturally distinct sample: people who judged the decision to cause harm as *unjustified* judged the harm as *intentional*, whereas people who judged the decision as *justified* judged the harm as *not intentional*. Crucially, the hypothesized pathway between the

different conditions and intentionality judgements was also supported—it is how the increase in profit affected the perceived justifiability of causing harm that affected intentionality judgements.

Experiment 3a

Experiment 3a tested the effect of justifiability in a similar way to Experiment 2, only instead of varying the justifiability of the harm by changing the amount of profit, this time we varied the costs (harm) and kept profits the same. Participants read a short vignette describing an event in which the Chief of Operations makes various changes to a movie based on a book. The reason for making the changes is constant across conditions, however, the costs of making the decision varied across conditions. Depending on which condition the participants were assigned to, the Chief of Operations either made very small changes and annoyed very few fans (Low Cost), a moderate amount of changes and annoyed some fans (Moderate Cost), or a large amount of changes and annoyed almost all fans (High Cost). We predicted that as the costs increased, people would judge the Chief of Operations' decision as less justified and perceive the harm as more intentional and, crucially, that the relationship between increased costs and increase harmful intentions would be mediated by reduced justifiability.

Method

Preregistration https://aspredicted.org/blind.php?x=cc2yh7

Procedure

Expect for recruiting American participants from Prolific, the procedure was the same as Experiment 2b.

Participants

As preregistered, we requested 300 American participants from Prolific. Prolific sent 329 and 43 were excluded for failing an attention check, leaving a total sample of 286 participants (M_{age} = 30.6, SD = 10.10). As with both previous studies, the majority were female (141 Female, 134 Male, 11 Gender diverse) and white (65.4% White, 8.0% Black, 16.8% Asian, 9.4% Other).

Measures

The measures were the same as in the previous experiment but with slight changes in wording to fit the vignette (e.g., “Did the Chief of Operations intentionally annoy the fans?”).

Vignette

> A new film was being made based on a popular book. During the production, the vice production manager said to the Chief of Operations “We are thinking of making some changes to the film. It will slightly reduce production costs but cause the film to be a tiny bit different from the book, somewhat annoying a small number of the book's hard-core fans / It will reduce production costs but cause the film to be moderately different from the book, annoying some fans of the book / It will reduce production costs but cause the film to be majorly different from the book, seriously annoying almost all fans of the book.
>
> The Chief of Operations decided to make the changes. Sure enough, a small number of hard-core fans were somewhat annoyed by the changes/ some fans of the book were annoyed by the changes/ almost all fans of the book were seriously annoyed by the changes.

Results

Key Results

Perceived Justifiability of the Decision. As predicted, a one-way ANOVA revealed a significant main effect of cost on the perceived justifiability of the decision $F(2,283) = 14.5$, $p < .001$, $\eta^2 = .093$). Also as predicted, a linear contrast showed a significant linear trend, $B = 0.70$, $p < .001$, indicating that as the costs increased, people became less likely to judge the decision as justified (Figure 6, Table 4).

Table 4

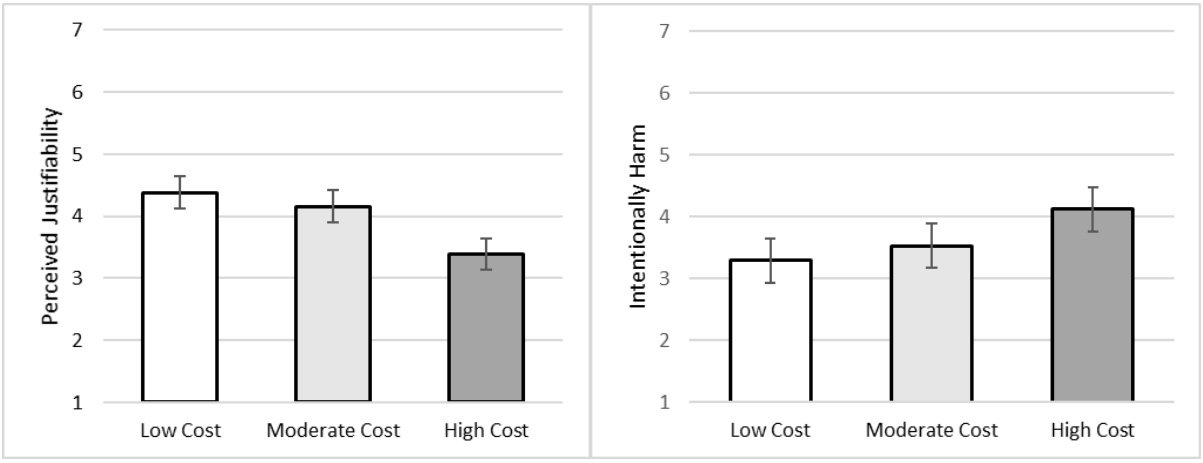
Means (and standard deviations) in Experiment 3a, split by condition

	Low Cost	Moderate Cost	High Cost
Perceived justifiability	4.38(1.22)[a]	4.16(1.27)[a]	3.39(1.48)[b]
Intentionally Harm	3.29(1.76)[a]	3.53(1.78)[ab]	4.12(1.86)[b]
Intentionally make changes	5.83(1.36)[a]	6.00(1.30)[a]	5.98(1.55)[a]

Note. In rows, means with different superscripts are significantly different ($p < .05$). Differences in means were tested using Bonferroni corrected planned comparisons.

Figure 6

Mean Perceived Justifiability and Intentionality Responses in Experiment 3a by Condition



Note. Perceived intentions and justifiability were measured on a 7-point Likert scale where 1 = Definitely Not, 4 = Maybe, and 7 = Definitely Yes. Error Bars = 95% CI

Perceived Intentions to Harm. As predicted, a one-way ANOVA revealed a significant main effect of cost on peoples' judgements of the Chief of Operation's intentions to annoy the fans, $F(2,283) = 5.42$, $p = .005$, $\eta^2 = .037$). Also as predicted, a linear contrast showed that there was a significant linear trend $B = -0.59$, $p = .002$, indicating that as the

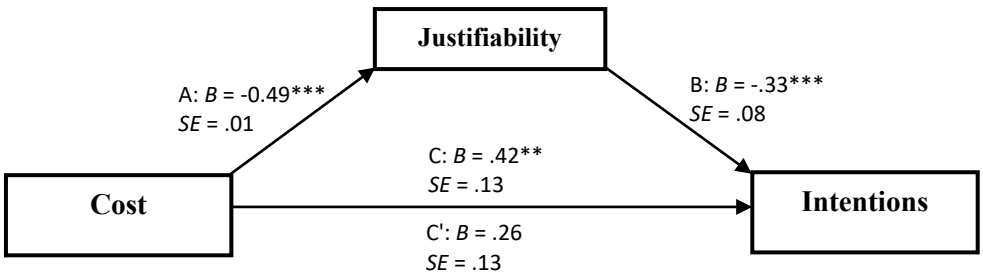
costs of the decision increased (i.e., amount of fans annoyed increased) people judged the harm as more intentional (Figure 6).

Perceived Intentions to make the Changes. As predicted, a one-way ANOVA showed that how costly the decision was had no effect on perceptions of the Chief of Operations intentionally making the changes, $F(2,283) = 0.417$, $p = .659$, $\eta^2 = .003$).

Pathway from Profit to Perceived Intentions to Harm via Perceived Justifiability. Mediation using the MEDMOD package in JAMOVI and 5000 bootstrap samples revealed that increases in the costs of the Chief of Operations' decisions was associated with increased perceived intentions to annoy the fans and that a decrease in the perceived justifiability of the decision mediated this link (indirect effect: 0.16, 95%CI:[0.05, 0.27], SE = 0.06, $p = .004$; Figure 7). Increase in cost was associated with decreased perceptions that the decision was justified ($B = -0.49$, 95%CI:[-0.69, -0.30], SE = 0.10, $p < .001$). An increase in the perceived justifiability of the decision was associated with a decrease in perceptions that the Chief of Operations intentionally annoyed the fans ($B = -0.33$, 95%CI:[-0.50, -0.15], SE = 0.09, $p < .001$). The remaining direct effect was slightly above the cut-off point ($B = 0.26$, 95%CI:[-0.01, 0.52], SE = 0.13, $p = .055$), indicating that the effect was fully mediated.

Figure 7

Mediation Model Depicting the Pathway from Increased Costs to Increased Perceived Intentions to Harm via Reduced Perceived Justifiability in Experiment 3a



Discussion

Experiment 3a extended the results from the previous experiments by showing that information about the justifiability of a person's decision—specifically its costs—influence people's judgements of whether the harm resulting from the decision was intentional: people judged that the Chief of Operations intentionally annoyed fans of the book when they thought the decision was unjustified (i.e., the costs outweighed the benefits), but not when they thought the decision was justified (i.e., benefits outweighed costs). Experiment 3a thus provides further support for the hypothesis that people's judgements of others' intentions are sensitive to information that justifies their decision to cause harm.

Experiment 3b

Experiment 3b aimed to test the generalizability and replicability of Experiment 3a with a new, culturally distinct sample (Polish compared to American). All stimuli and predictions were the same as Experiment 3a.

Method

Preregistration https://aspredicted.org/blind.php?x=g6m3gf

Procedure

Expect for recruiting Polish nationals fluent in English from Prolific, the procedure was the same as Experiment 3a.

Participants

As preregistered, we requested 300 Polish participants fluent in English from Prolific. Prolific sent 325 and 50 were excluded for failing an attention check, leaving a total sample of 275 participants. In contrast to the previous study, the sample was an average 7 years

younger (M_{age} = 23.6, SD = 6.59), almost all participants identified as white (99.6% White, 0.4% Black), and the majority were male (195 Male, 77 Female, 3 Gender diverse).

Measures

The measures were the same as the previous experiment.

Vignette

The vignettes were the same as in Experiment 3a.[2]

Results

Key Results

Perceived Justifiability of the Decision. As predicted, a one-way ANOVA revealed a significant main effect of cost on the perceived justifiability of the decision $F(2,272) = 5.83$, $p = .003$, $\eta^2 = .041$). Also as predicted, a linear contrast showed a significant linear trend, $B = 0.48$, $p < .001$, indicating that as the costs increased, people became less likely to judge the decision as justified (Figure 8, Table 5).

Table 5

Means (and standard deviations) in Experiment 3b, split by condition

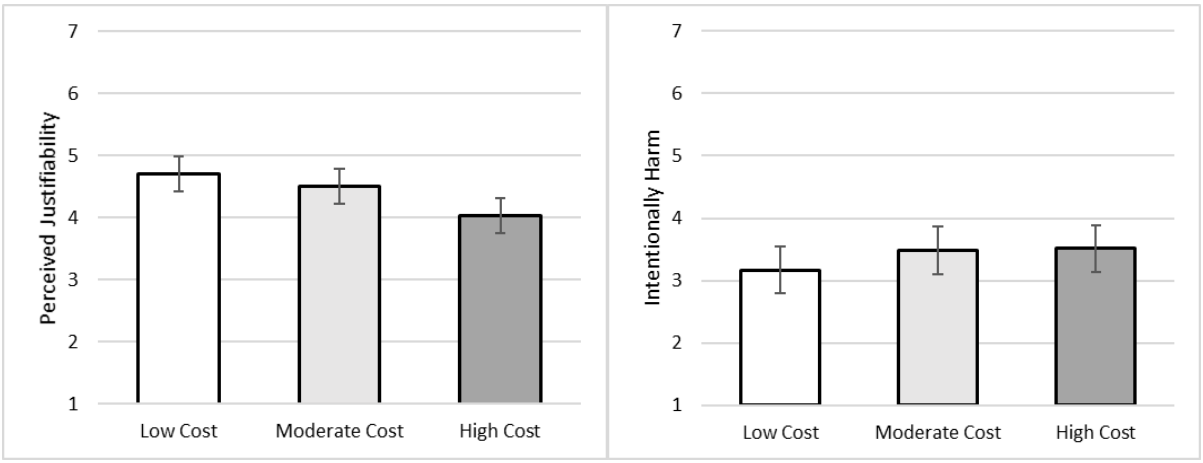
	Low Cost	Moderate Cost	High Cost
Perceived justifiability	4.71(1.20)[a]	4.51(1.40)[a]	4.03(1.48)[b]
Intentionally Harm	3.17(1.74) [a]	3.48(1.86) [a]	3.51(1.96) [a]
Intentionally make changes	5.22(1.60) [a]	5.39(1.59) [a]	5.40(1.59) [a]

Note. In rows, means with different superscripts are significantly different ($p < .05$). Differences in means were tested using Bonferroni corrected planned comparisons.

[2] To ensure the vignette was appropriate for use with English speaking Polish nationals we consulted with one of the authors Polish friends.

Figure 8

Mean Perceived Justifiability and Intentionality Responses in Experiment 3b by Condition



Note. Perceived intentions and justifiability were measured on a 7-point Likert scale where 1 = Definitely Not, 4 = Maybe, and 7 = Definitely Yes. Error Bars = 95% CI

Perceived Intentions to Harm. Potentially due to the manipulation having a smaller effect on justifiability compared to the Experiment 3a ($\eta^2 = .041$ vs $\eta^2 = .093$, a difference of a small effect size, $d = 0.23$. More on this below.), we failed to detect an effect of cost on people's judgements of the Chief of Operations intentions to annoy the fans, $F(2,272) = 0.938$, $p = .393$, $\eta^2 = .007$). Critically, the number of people recruited was based on the medium effect sizes detected in previous experiments, and, given the weaker effect of the manipulation on perceived justifiability, the reason we may not have detected an effect is due to insufficient power. Indeed, Post-hoc power analysis using G*Power (Faul et al., 2009) indicted that we only had 22% power to detect the effect. Nonetheless, while no significant trend was detected ($p = .219$), the pattern of means was still in the direction predicted (Low < Moderate < High, Table 5).

Perceived Intentions to make the Changes. As predicted, a one-way ANOVA showed that how costly the decision was had no effect on perceptions of the Chief of Operations' intentions to make the changes, $F(2,272) = 0.395$, $p = .674$, $\eta^2 = .003$) (Table 5).

Pathway from Profit to Perceived Intentions to Harm via Perceived Justifiability. Mediation using the MEDMOD package in JAMOVI and 5000 bootstrap samples revealed that increases in the costs of the Chief of Operations' decisions was associated with increased perceived intentions to annoy the fans and that a decrease in the perceived justifiability of the decision mediated this link (indirect effect: 0.10, 95%CI:[0.02, 0.18], SE = 0.04, p = .018). Increase in cost was associated with decreased perceptions that the decision was justified (B = -0.34, 95%CI:[-0.53, -0.14], SE = 0.10, p < .001). An increase in the perceived justifiability of the decision was associated with a decrease in perceptions that the Chief of Operations intentionally annoyed the fans (B = -0.30, 95%CI:[-0.45, -0.14], SE = 0.08, p < .001). The remaining direct effect was non-significant (B = 0.07, 95%CI:[-0.20, 0.34], SE = 0.14, p = .569), indicating that the effect was fully mediated.

Exploratory Analyses

People who judged harm as justified judged it as not intentional. Exploratory analyses indicate that TJM's key prediction that people who judge a decision to cause harm as *unjustified* will judge the harm as *intentional*, whereas people who judge the decision as *justified* will judge the harm as *not intentional* was supported: people who judged the decision as definitely unjustified (i.e., reported justifiability of 1, N = 9) judged the harm as intentional (M = 4.89, SD = 2.15), whereas people who thought the decision was definitely justified (i.e., reported justifiability of 7, N = 14) judged the harm as not intentional (M = 2.57, SD = 1.83).

Comparison of Experiment 3a and 3b. Overall, the same key pattern was present in Experiment 3b as 3a. However, there did appear to be some *prima-facie* differences (e.g., people in 3b seemed to judge the harm as more justified). We wanted to investigate potential

causes of this further, so we combined the data from both experiments and ran some exploratory analyses.

A 2x3 ANOVA (experiment x condition) revealed a significant effect of condition, $F(2,555) = 19.10$, $p < .001$, $\eta^2 = .063$, and Experiment, $F(2,555) = 15.10$, $p < .001$, $\eta^2 = .025$, on perceived justifiability (the interaction was not significant $p = .466$). The same pattern that was present in each experiment emerged: as the costs increased, people became less likely to judge the decision as justified, $B = 0.59$, $p < .001$. Curiously, people in Experiment 3b were more likely to judge the decision as justified ($M = 4.43$, $SD = 1.39$) than in 3a ($M = 3.97$, $SD = 1.39$), $t(555) = 3.90$, $p_{bonferroni} < .001$, $d = 0.165$. As age is a predictor of perceived justifiability in loose countries (Jiang et al., 2015), this difference could be attributed to the fact that the sample in 3b was significantly younger than 2a ($M_{DIFF} = 7$ years), $t(559) = 9.74$, $p < .001$, $d = 0.82$. Indeed, follow-up regression analysis showed that age was a significant predictor of perceived justifiability ($B = -.01$, 95%CI:[-.027, -.002], $t(560) = -2.32$, $p = .021$).

We ran a second 2x3 ANOVA (experiment x condition) testing for any differences in intentionality judgements across experiments and, crucially, if there was an effect of cost on intention judgements when the data from the two Experiments were combined. As expected, there was a main effect of cost on perceived harmful intentions, $F(2,555) = 4.78$, $p = .009$, $\eta^2 = .017$, such that as the costs increased so too did perceived harmful intentions, $B = -0.41$, $p = 002$. No effect of experiment on judgements of intentions to harm was detected, $F(2,555) = 2.86$, $p = .091$, $\eta^2 = .005$. Collating across both experiments, TJM's predictions were supported: as the costs increased, perceived justifiability would decrease, and, consequently, perceived harmful intentions increased.

Discussion

Experiment 3b tested the generalizability and replicability of 3a by keeping the stimuli constant and using a culturally distinct sample (Polish compared to American). Overall, Experiment 3b replicated the key results from 3a: judgements of harmful intentions were mediated by the perceived justifiability of the harm. When people judged the decision as unjustified they judged the harm as intentional, but when they judged the decision as justified the harm was judged as not intentional.

Experiment 3b had limitations not present in 3a. For example, some people thought that the Vice-Production manager was the boss of the Chief of Operations (CoO) and was pressuring the CoO into making the changes, thus, justifying the CoO's decision regardless of condition (e.g., "He was made to do it by his boss…"). This, combined with the fact that the sample was significantly younger and therefore more likely to judge the decision as justified, were presumably the causes for the weaker manipulation and for not being able to detect an effect of cost on intentionality judgements.

Combined, however, the results of Experiment 3a and 3b, support TJM's predictions—people's judgements of whether the harm was intentional tracked the perceived justifiability of the harm. Manipulating the justifiability of causing harm has now been shown to affect attributions of intentions in people from 4 different countries (UK, SA, USA, and Poland) with 3 separate vignettes—one manipulating costs, the others benefits. Experiment 4 sought to test the effect of justifiability on intention judgements using a different paradigm.

Experiment 4

Experiment 4 aimed to test whether new information about the justifiability of causing a harm is incorporated into people's judgements or if people are anchoring on their initial judgement and not changing their attributions. Similar to Monroe and Malle's (2017, 2019)

updating blame paradigm, people will read a scenario about a person who causes harm and make an initial judgement about the person's intentions and the justifiability of the harm. After making this initial judgement people will be provided with new information that increases, decreases, or does not affect the justifiability of the harm and given an opportunity to update their judgements. TJM predicts that people will update their judgements of a person's intentions based on new (un)justifying social information. Specifically, that judgements that the harm was intentional will decrease when people are given new justifying social information. Conversely, that judgements that the harm was intentional will increase when people are given new unjustifying social information. Furthermore, that new information that does not affect the justifiability of the harm (irrelevant social information) will not affect people's judgements.

Method

Preregistration https://aspredicted.org/blind.php?x=yr8qi8

Procedure

Participants read 15 different vignettes, presented in random order, about agents who caused harm. After reading each vignette, participants made an initial judgement about whether they thought the harm was intentional and whether they thought the harm was justified. After making their initial judgement, participants were given more information about the scenario. There were three different types of new information provided: Justifying social information, which makes the harm appear more justified; Unjustifying social information, which makes the harm appear less justified; and irrelevant information, which does not affect the justifiability of the harm. Each new type of information was presented five times (i.e., participants got new unjustifying information five times, new justifying social information five times, and new irrelevant information five times). After reading the new

information participants were shown their initial judgements on a new screen and given the opportunity to update their judgements with the prompt, “If your judgements about … have changed, please indicate by moving the sliders from their original position on the next screen.”

Participants

In within-subjects design studies with three conditions and five stimulus replication per design cell, G-power (Faul et al., 2009) recommends using a minimum sample size of 33 participants to detect a moderate effect size ($\eta 2 = .06$) with .8 power. We will request a sample of 50 students to ensure that we have sufficient power (> .8) even after excluding inattentive participants.

50 undergraduate psychology students from the University of Canterbury, New Zealand were requested to participate in this study for course credit. Seven were excluded for failing an attention check, leaving a final sample of 43 participants (M_{age} = 24.0, SD = 7.70). The majority were female (30 Female, 13 Male) and of New Zealand European ethnicity (72% New Zealand European, Other 16%, 5% Māori, 5% Chinese).

Measures

The measures were the same as in previous studies but with slight changes in wording to fit the vignettes (e.g., “Did Andrew intentionally cause the old lady to get hurt? Was Andrew justified in not giving up his seat?”).

Example Vignette

> A little old lady came onto the bus and needed a place to sit. Andrew knew she’d stumble and hurt herself as the bus took off if she didn’t get a seat. Yet, Andrew did

not offer his seat to her. She didn't get a seat and, sure enough, as the bus took off the lady stumbled and hurt herself.

Example New Information

Andrew had recently broken his knee and standing on the bus would have caused him great pain.

Results

Results are robust to exclusions—the same pattern of results is shown with and without exclusions. As we had preregistered excluding participants, the results we report here are with participants excluded.

Manipulation Check

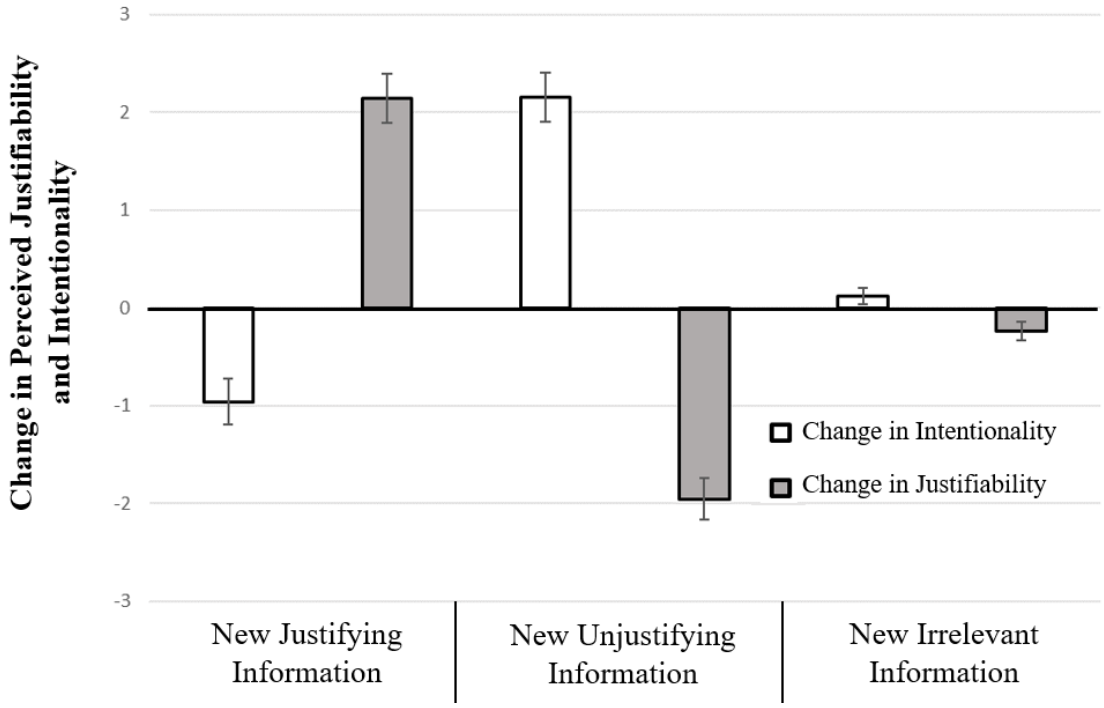
As preregistered, we went through each of the vignettes and checked that the changes in mean justifiability ratings were in the predicted direction for each of the new type of social information. All changes in means were in the predicted direction, thus, analyses reported here collapse across vignettes with the same type of new social information.

Key Results

The Trade-off Justification Model predicts that intentionality judgements will decrease or increase as a function of the agent's perceived justifiability in causing harm: decreased justifiability increases intentionality, whereas increased justifiability decreases intentionality. A within-subjects ANOVA showed that new type of information presented accounted for 43% of the variance updated intentionality judgements, $F(2, 642) = 245.9$, $p < .001$, partial $\eta^2 = .434$ (Figure 9). Crucially, the preregistered post-hoc tests were all supported (discussed in detail in the following three sub-sections).

Figure 9

Mean Change in Intentionality and Justifiability Judgements as a Function of New Information Type



Note. Change scores = *Updated Judgement – Initial Judgement*. Error Bars = 95% CI.

New Justifying Social Information Increased Perceived Justifiability of Harming and Decreased Perceived Intention to Harm. As predicted, when participants were provided with new justifying social information they updated their judgements of the justifiability of harming, viewing it as more justified (M_{CHANGE} = 2.14, SD_{CHANGE} = 1.83, Table 6, Figure 9). Crucial to TJM's prediction, people also updated their judgements of the person's intentions, perceiving the harm as less intentional (M_{CHANGE} = -0.95, SD_{CHANGE} = 0.96).

Table 6

Means (and standard deviations) in Experiment 4 by Type of New Information

	Intentionality					Justifiability				
Condition	**Initial**	**Updated**	***t(642)***	p_{Tukey}	***d***	**Initial**	**Updated**	***t(642)***	p_{Tukey}	***d***
New Justifying Information	4.48(1.76)	3.53(1.97)	9.51	<.001	-0.57	2.92(1.46)	5.07(1.77)	-21.69	<.001	1.17
New Unjustifying Information	2.73(1.54)	4.88(1.87)	-21.37	<.001	1.18	4.38(1.48)	2.42(1.49)	19.76	<.001	-1.25
New Irrelevant Information	4.70(1.62)	4.82(1.63)	-1.2	0.837	0.19	2.62(1.42)	2.38(1.42)	2.40	0.158	-0.33

New Unjustifying Social Information Decreased Perceived Justifiability of Harming and Increased Perceived Intention to Harm. As predicted, when participants were provided with new unjustifying social information they updated their judgements of the justifiability of harming, viewing it as less justified (M_{CHANGE} = -1.95, SD_{CHANGE} = 1.56, Table 6). Consequently, people also updated their judgements of the person's intentions, perceiving the harm as more intentional (M_{CHANGE} = 2.15, SD_{CHANGE} = 1.82).

New Irrelevant Social Information Had no Effect on Perceived Intention to Harm. As predicted, when participants were provided with new irrelevant social information they did not change their judgements of the justifiability of harming (M_{CHANGE} = 0.24, SD_{CHANGE} = 0.72, Table 6). Crucially, people also did not change their judgements of the person's intentions—the harm was perceived as no more or less intentional (M_{CHANGE} = 0.12, SD_{CHANGE} = 0.63).

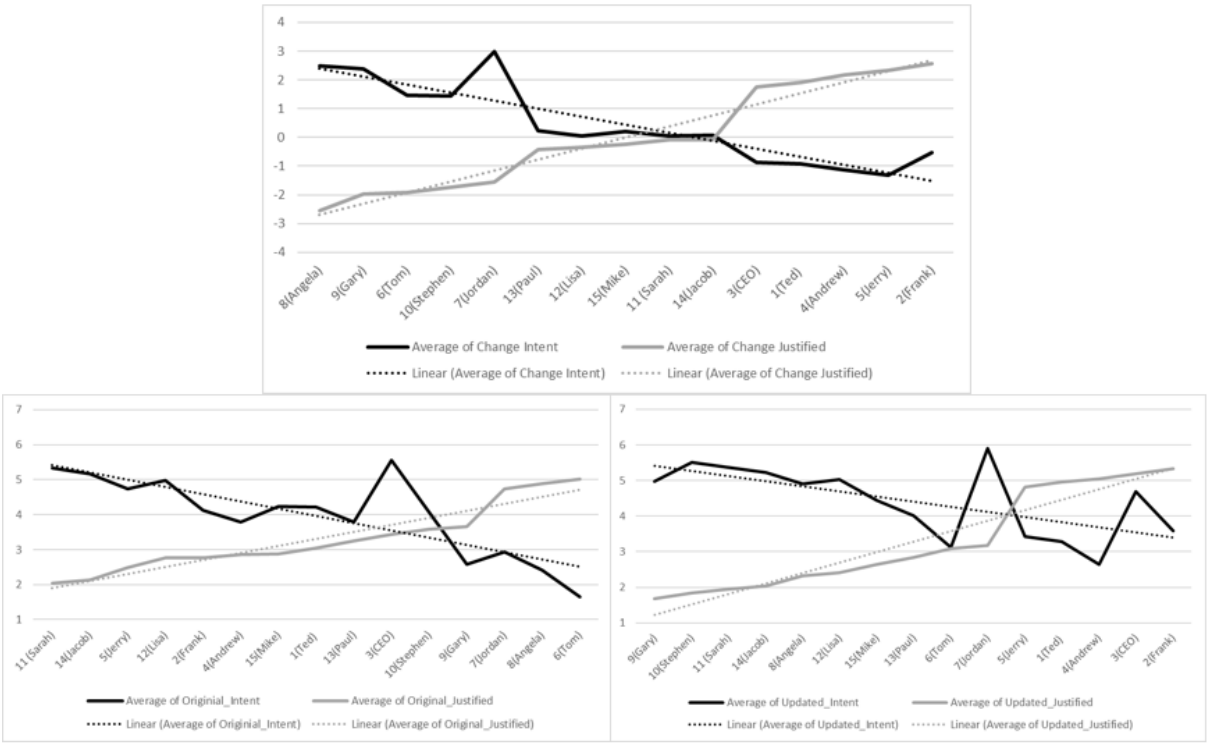
Exploratory Analyses

In Experiment 1, when looking across vignettes, the ones that were perceived as less justified were judged as more intentional. We were interested if this trend replicated. Thus, as in Experiment 1, we used linear mixed models in Jamovi (clustering by vignette and including random slopes and intercepts for each vignette) to estimate the effect of justifying

social information on intentions at the vignette level. Results show the same overall pattern as in Experiment 3 for the original judgements (B = -0.32, 95%CI:[-.41, -.24], t(11.6) = - 7.18, p < .001, $R^2_{conditional}$ = .35, Figure 10 bottom left), updated judgements (B = -0.42, 95%CI:[-.50, -.35], t(13.9) = - 6.98, p < .001, $R^2_{conditional}$ = .36, Figure 10 bottom right), and change judgements (B = -0.37, 95%CI:[-.47, -.28], t(10.5) = -7.81, p < .001, $R^2_{conditional}$ = .45, Figure 10 bottom right).

Figure 10

Line Graph showing the Negative Association between Mean Intentionality and Mean Justifiability Responses across the Series of Vignettes in Experiment 4



Note. Bottom left shows the relationship between justifiability and intentionality for participant's initial judgements, bottom right shows the relationship between justifiability and intentionality for participant's updated judgements, and the top figure shows the relationship between justifiability and intentionality for participant's change scores. Change scores = *Updated Judgement – Initial Judgement.* The full vignettes are available in the appendix.

Vignette number 7 (Jordan) was a notable exception to the trend. Presumably, this is because the vignette included an extra-piece of social information—hurtful words towards the victim—that increased certainty that the harm was intentional. Indeed, common open-

ended responses such as, “Jordan's coworker responded in a very unkind way when she asked why she wasn't invite to lunch. This suggests that Jordan's coworkers intentionally did not invite [her] to lunch despite knowing that it would upset her” and “...the co worker didn't need to answer her in a harsh way, that shows that they had done it [(hurt her)] intentionally.” support this interpretation.

Discussion

Experiment 4 extended support for TJM by showing that people’s judgements of whether someone intentionally caused harm tracks information about the justifiability of causing the harm: when people were given new information that decreased the justifiability of causing the harm, they changed their judgements, perceiving the harm as more intentional. Moreover, when people were provided with new information that increased the justifiability of causing the harm they also changed their judgements, perceiving the harm as less intentional. Crucially, however, when people were provided with new information that did not affect the justifiability of the harm, people did not change their judgements.

Experiment 4 also showed the same overall pattern as in Experiment 1: across a wide range of scenarios a general trend emerged showing that harms judged as more justified are perceived as less intentional. There was the exception of vignette number 7 (Jordan) which was judged as far more intentional than other scenarios with similar justifiability. This is likely due to this vignette including an extra piece of social information—hurtful words towards the victim. This result does not conflict with TJM as TJM argues that people utilize various pieces of social information (but particularly information about the justifiability of the harm) when inferring others’ intentions. Thus, it makes sense that when more information is provided, people attribute more intentionality to the person.

We argue this and the previous experiments have validated the hypothesized mechanism behind how people attribute intentions. We now aim to build off the finding that justifiability is linked to intentionality judgements by investigating how observer's dispositional beliefs can affect the perceived justifiability of a harm, causing some people to perceive a harm as more or less intentional than others.

Experiment 5

The Trade-off Justification Model hypothesizes that differences in whether causing harm is perceived as intentional is due to differences in the perceived justification of causing the harm—unjustified harms are judged intentional—but how are observer's judging the justifiability of the harm? Put differently, what causes observers to arrive at different judgements of the justifiability of a harm, and, consequently, whether a harm is judged intentional? Previous research has shown that what actions are perceived as unjustified differs from culture to culture (Bahník et al., 2019; Haerpfer et al., 2020; Jiang et al., 2015). Experiment 5 thus builds on the previous findings by adding a cross-cultural component. We reason that people's cultural backgrounds affect perceptions of the justifiability of certain actions and therefore how intentional the actions' harmful side-effects are perceived to be.

In the Chinese cultural context education is highly prized—it is perceived as key to being successful and children spend many hours in afterschool education programs, known as cram schools (Larmer, 2014, December 31). In the USA context, education is comparatively less prioritized and attending cram schools is far from the norm. We utilize this difference and created a scenario (see below) where a father sends their son to an afterschool exam preparation program/programme (cram school) with the harmful consequence of greatly upsetting them. As education is highly prized and sending children to cram schools is normal in China, we predict that Chinese participants would judge the father's decision as more justified and the harm as less intentional than American participants.

The harm in this experiment is minor compared to typical harms in the literature (upsetting a child vs destroying the environment). We kept the harm minor to keep as much consistent across the vignettes as possible while still having the harm question make sense to people from both nationalities. If we had based our scenario off a taboo that existed in one country but not another (eating meat during certain periods of the year, for example) then the question about whether someone intentionally caused harm by deciding to violate the taboo would be nonsensical in one cultural context because no harm (taboo violation) would be perceived.

Method

Preregistration https://aspredicted.org/blind.php?x=iu8tj6

Procedure

Participants were assigned to one of two conditions based on their nationality. The changes between conditions were made to reflect the different cultural contexts (e.g., different spellings and names) and are non-substantial. Participants read a vignette (see below) and then answered questions about the justifiability of the Father's decision and the Father's intentions. The order of the justifiability and intention questions were randomized to avoid possible order effects. Participants then completed demographics (gender, age, ethnicity) and were debriefed.

Participants

As preregistered, we requested 100 non-Asian American Nationals and 100 Chinese nationals who are fluent in English from Prolific. Prolific sent 202 participants. Six requested their data be deleted, 13 were excluded for failing the attention check, and 7 more were excluded for having a nationality that conflicted with the sampling criteria (e.g., we excluded participants who indicated they had American or Canadian nationality when we requested for

Chinese nationals). Leaving a final sample of 182 participants (88 Chinese nationals, 94 US nationals). Chinese nationals were slightly younger (M_{age} = 27.90, SD = 7.58, vs M_{age} = 30.90, SD = 11.00) and had more female participants (54 Female, 34 Male vs 37 Female, 54 Male, and 3 Gender diverse).

Measures

The measures were the same as in previous studies but with slight changes in wording to fit the vignettes (e.g., “Did Jordan's/Jun's Father intentionally upset him?”).

Vignette

> Jordan's/Jun's mother said to his father, "should we enroll/enrol Jordan/Jun in the afterschool exam preparation program/programme (cram school)? It will improve his grades from B+ to straight A's but will also mean he won't have time to play soccer anymore. He'll be really upset at having to quit the team. What do you think?" Jordan's/Jun's father replied, "He must focus more on his education than playing games. I'm enrolling him." Jordan/Jun was sent to the afterschool exam preparation program/programme (cram school). Sure enough, Jordan/Jun was very upset at having to quit playing soccer.[3]

Results

As predicted, Chinese nationals rated the father’s decision to send their child to cram school as more justified and the decision’s harmful consequence as less intentional than American nationals (Table 7).

[3] When designing this scenario we consulted with one of the authors Chinese friends to ensure it was framed appropriatly.

Table 7

Means (and standard deviations) in Experiment 5 by Nationality

Measure	Nationality	*M*(*SD*)	*t*(180)	*p*	*D*	*d* 95%CI
Justified	American	3.61(1.55)	3.69	< .001	0.55	[0.24, 0.85]
	Chinese	4.43(1.42)				
Intentionality	American	3.84(1.80)	-4.32	< .001	-0.63	[-0.95, -0.33]
	Chinese	2.71(1.78)				

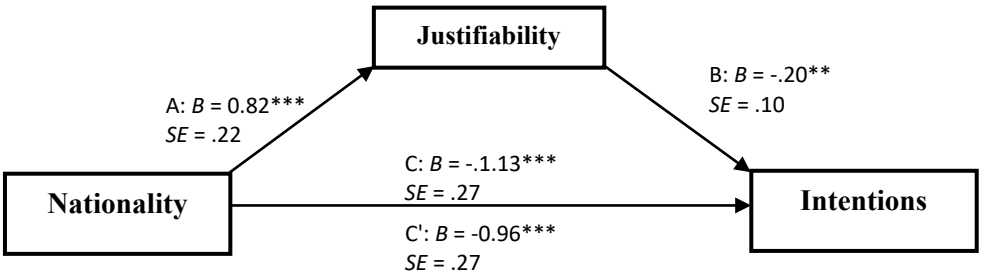
Pathway from Nationality to Perceived Intentions to Harm via Perceived Justifiability. Mediation using the MEDMOD package in JAMOVI and 5000 bootstrap samples revealed that Chinese nationals were associated with a decrease in ratings that the Father intentionally upset their child, however, the indirect effect of an increase in the perceived justifiability of the father's decision to send their child to cram school did not mediated this link (indirect effect: -.17, 95%CI:[-0.37, 0.00], SE = 0.09, p = .068; Figure 11). Though the model did not reach significance when bootstrapped, the non-bootstrapped model did (p = .049) and the coefficients were comparable to the previous models. We thus reason that the effect was mediated but lacked power to reach significance. Indeed, we based power off the t-test, not mediation. Had we gone for mediation, the recommended sample to reach sufficient power for this model is around 400 (Fritz & MacKinnon, 2007), over double what we recruited.

Though the overall bootstrapped model was not significant, the Chinese national sample was still associated with increased perceptions that the father's decision was justified (B = 0.82, 95%CI:[0.39, 1.26], SE = 0.22, p < .001), and an increase in the perceived justifiability of the father's decision was still associated with a decrease in perceptions that the father intentionally upset their son (B = -0.20, 95%CI:[-0.40, -0.00], SE = 0.10, p = .042). There was a large, significant remaining direct effect (B = -0.96, 95%CI:[-1.65, -0.61], SE = 0.27, p < .001) indicating that the effect was not fully mediated (Figure 11). Overall, the

model supports the idea that cultural differences in the perceived justifiability of causing harm leads to differences in whether the harm is perceived as intentional.

Figure 11

Mediation Model Depicting the Pathway from Nationality to Reduced Perceived Intentions to Harm via Increased Perceived Justifiability in Experiment 5



Discussion

Previous studies on TJM have not specifically tested where justifiability comes from. While previous research has shown that roles (Rowe et al., 2020) and taboos (Vonasch & Baumeister, 2017) affect perceived justifiability, the interaction between the observer and the scenario has not been tested. This study shows that culture sets the norms/values people use to evaluate the justifiability of causing harm and, thereby, the intentionality of different acts. Experiment 5 thus supports an important extension of the hypothesized link between justifiability and intentionality: people from different nationalities judged the intentionality of a harm differently based on how their cultural background affected the perceived justifiability of the harm. Chinese participants viewed the decision to send a child to cram school as more justified and the resulting harm (the child becoming upset) as less intentional than American participants.

Because the harm had to be minor so that the question made sense to people from both nationalities it was viewed as at least somewhat justified in both countries. Thus, the results were limited to showing that one culture rated the harm as *less* intentional than

another. However, we believe that stronger effects—such as a harm being judged intentional in one culture but not intentional in another—can in principle be shown.

Experiment 6

The previous experiment tested how culture can set the values and norms people use to evaluate justifiability and, consequently, intentionality. In this study we examine whether a specific group affiliation/identity can have the same effect.

People identify with all types of groups—some people identify more with authoritarian groups, others with libertarian groups; some with affirmative action, others with equal opportunity; some with Black Lives Matter, some with Blue Lives Matter. Presumably, these differences lead to variation in whether certain actions—such as the use of force by officers—are justified, and, consequently, judgements of whether resulting harms were intentional. Experiment 6 thus investigates the effects of people's group identities—specifically, their overall ideological stance towards Black/Blue Lives Matter—on judgements of whether an officer was justified in their use of force and whether the officer is perceived to have intentionally caused a protestor (Martin Gugino) harm.

We have three key hypotheses:

> H1: People's level of identification with Black/Blue Lives Matter will predict their judgements of the justifiability of the officer's actions: People who identify more with Black Lives Matter will judge the officer's actions as more unjustified than people who identify more with Blue Lives Matter.
>
> H2: People's perceived justifiability of the officer's actions will predict their intentionality judgements: People who judge the push as unjustified will judge the harm more intentional.

H3: The link between people's overall level of identification with Black/Blue Lives Matter and judgements whether the harm was intentional will be mediated via perceived justifiability.

Method

Preregistration https://aspredicted.org/blind.php?x=fn3sh7

Participants

We recruited participants from Amazons Mechanical Turk on the 7th of July, 2020 less than 1 month after the incident when it was still a lively topic on social media. We requested 250 participants based on prior research indicating this is when correlations stabilize (Schönbrodt & Perugini, 2013). Mturk sent 261 participants; 40 were excluded from the analyses because they failed an attention check and 12 because they did not complete the study, leaving a total sample of 209 participants (M_{age} = 35.7, *SD* = 11.0) (137 Males, 71 Females, and 1 non-binary). The majority of participants were White (61%), Black or African American (21%), or Asian (14%), tended to identify more strongly with Black than Blue Lives Matter (M = 4.48, SD =1.58) (on a scale of 1-6 where 1 = more strongly identify with Blue Lives Matter and 6 = more strongly identify with Black Lives Matter), and tended to be more Liberal (*N*=107) compared to Conservative (*N*=63).

Procedure

All participants watched the same video of the officer pushing Martin Gugino, who then fell on the ground and head started bleeding (video available via OSF https://osf.io/u9sp6/?view_only=9cb9b9749dae4a34a72e27ae083c5d50). Participants answered all of the same questions, however, the order in which this was done was counterbalanced. Participants either answered the key 'Identify Black vs Blue Lives Matter' question and related demographic questions first, and then watched the video of the officer

pushing Martin Gugino and answered questions about the officer's intentions etc. Or, participants watched the video and answered the related questions about the officer's intentions etc., before answering the key 'Identify Black vs Blue Lives Matter' question and related demographic questions. The order of the dependent variables (perceived intentions, justifiability, blameworthiness) were counterbalanced, appeared one at a time, and participants could not go back and change their answers. The last two questions were always how much participants support Black (Blue) lives matter, presented in random order.

Measures

Responses to the measure of identification with Black vs Blue lives Matter was recorded on a 6-point Likert scale where 1 = identify most with Blue Lives Matter and 6 = identify most with Black Lives Matter. Intentionality and justifiability were measured on 7-point Likert scales where 1 = definitely not, and 7 = definitely yes, and blame was measured on a 7-point Likert scale where 1 = no blame at all, and 7 = the most blame you would ever give.

Level of identifying with Black vs Blue Lives Matter. "Currently in the U.S. there is a lot of debate over policing and race. These issues are complex and people sometimes find themselves identifying with different movements. We want you to think about your identification with two different movements: the Black lives Matter movement and the Blue lives Matter movement. Of these two movements, which do you more strongly identify with?".

Intentions to Harm. "Did the Officer intentionally harm the man?"

Intentions to Push. "Did the Officer intentionally push the man?"

Justifiability. "Do you think the police officer was justified in pushing the man?"

Blame. "How much blame does the officer deserve?"

Attention check. "Explain why you think the officer was/was not justified in pushing the man." Participants who gave off-subject reasoning were excluded (e.g., "Nice").

Measurement Validity Check.

The validity of the 'identify' question was checked by asking on separate 6-point Likert scales where 1 = not at all and 6 = a great deal the following questions:

Support for Blue Lives Matter. "How much do you support Blue Lives Matter"

Support for Black Lives Matter. "How much do you support Black Lives Matter"

Results

Before analyzing the key results, we first investigated the validity of our key 'identify' measure and whether there were any order effects.

Measurement Validation

Consistency Check. As preregistered, the 'identify' question's validity was investigated by checking that the movement participants most strongly identified with was also the movement they most strongly supported. For example, if a person responded '5' to the 'identify' question (indicating they more strongly identified with the Black Lives Matter movement) they were judged to be consistent if they more strongly (or equally strongly) supported Black compared to Blue Lives Matter. Only 10 responses (5%) were inconsistent, indicating that the measure was valid. All key results were robust to excluding or including inconsistent responses. In an effort to be conservative with exclusions, we decided to report all analyses with these participants included.

Exploratory Measurement Check. Further support for the validity of the identify measure was obtained via assessing convergent validity. If the measure is valid, support for Blue Lives Matter should be negatively correlated with Identifying more with Black compared to Blue lives matter and support for Black Lives Matter should be positively correlated with Identifying more with Black compared to Blue Lives Matter. Indeed, this is what we found: Support for Blue Lives Matter was negatively correlated with identifying with Black compared to Blue Lives Matter (r = -.38, p < .001) indicating that the more strongly a person supported Blue Lives Matter the more strongly they identified with Blue compared to Black Lives Matter. In comparison, support for Black Lives Matter was very strongly positively correlated with the 'identify' question (r = .85, p < .001), indicating that the more strongly a person supported Black Lives Matter the more strongly they identified with Black compared to Blue Lives Matter.

Order effects

A One-Way ANOVA revealed that video presentation order (whether participants saw the demographic questions or video first) had no effect on any of the DVs or the key 'identify' question (all p's > .07). There was an effect of video presentation order on support for Blue Lives Matter ($M_{\text{VideoFirst}}$ = 2.85, SD = 1.66, $M_{\text{DemographicsFirst}}$ = 3.33, SD = 1.76, p = .045). People who saw the video first tended to support Blue Lives Matter less compared to those who answered the demographic questions first. However, as this question always came after the video, the effect is likely not attributable to its order of presentation. As no other effects were significant, video presentation order was collapsed into one group.

A separate One-Way ANOVA revealed no effect of question order on attributions of intentions, perceived justifiability, or blameworthiness (all p's > .06). Thus, responses were collapsed into one group.

An independent samples t-test revealed that the order the 'support' question had no effect on how much participants supported Black, $t(207) = 1.66$, $p = .098$, or Blue Lives Matter, $t(207) = .04$, $p = .968$.

Key results

Predicting Perceived Justifiability of the Push. We regressed participants' perceived justifiability onto level of identifying with Blue/Black Lives Matter (i.e., using degree of identifying with Blue/Black Lives Matter to predict perceived justifiability). As predicted, the more people identified with Black compared to Blue Lives Matter, the less they perceived the push to be justified, $B = -.59$, 95%CI:[-.59, -.34], $t(207) = 7.57$, $p < .001$, $R^2 = .217$. Thus, people's overall ideological stance towards Black/Blue Lives Matter was consistent with their judgements of a relevant police action. Supporters of Black Lives Matter were more likely to judge a violent act by police as unjustified (presumably, because they perceived the act to be violent, unnecessary, or undeserved), whereas supporters of Blue Lives Matter were more likely to judge the same act as justified (presumably, to preserve order and control over the situation).

Participants' open-ended responses provided some support that these were the reasons some people thought the push was or was not justified. For example, 11 people who identified more strongly with Black Lives Matter included in their explanations key words such as "excessive", "unnecessary", or "didn't deserve" indicating that they thought the officer's violent action was unjustified because it was unnecessary, excessive, or that the man did not deserve to be treated like that. Comparatively, only 1 person who identified more with Blue Lives Matter more included these key words.

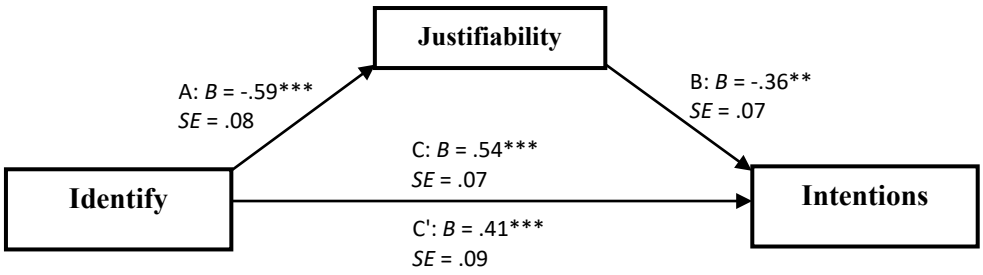
Predicting Perceived Intentions to Harm. In a separate regression, participants' ratings of the officer's intentions to harm the man were regressed onto perceived justification (i.e., using perceived justification to predict perceived intentions). As hypothesized, the more people judged the officer's actions as justified, the less they perceived the harm to be intentional (B = -.372, 95%CI:[-.53, -.28], $t(207) = 6.36$, $p < .001$, $R^2 = .164$). Thus, as predicted by TJM, people's judgements about the justifiability of the officer's actions were linked to their intention judgments: people who judged the action as unjustified thought it was more likely to be intentional, whereas people who judged the action as justified thought it was more likely to be unintentional.

Predicted Pathway to Perceived Intentions: Ideological Stance to Perceived Intentions via Perceived Justifiability. As predicted, mediation using the MEDMOD package in JAMOVI and 5000 bootstrap samples revealed that strength of identifying with Black compared to Blue Lives Matter increased perceived harmful intentions via a decrease in perceived justifiability(indirect effect: 0.131, SE = 0.05, 95%CI:[.04, .23], p = .007; Figure 12). Increased strength in identifying with Black compared to Blue Lives Matter decreased perceptions that the officer's actions were justified (B = -.59, 95%CI:[-.75, -.44], SE = .08, p < .001). A decrease in perceived justifiability was associated with an increase in perceived harmful intentions (B = -.22, 95%CI:[-.36, .08], SE = .07, p = .002). After accounting for the indirect effect, the direct effect was still significant (B = .41, 95%CI:[.24, .58], SE = .09, p <.001), meaning perceived justifiability partially mediated the relationship between ideology and perceived intentions. Thus, as hypothesized, peoples ideological stance towards Black/Blue Lives Matter affected their interpretation of an officer's intentions because of how their ideological stance affected their judgments of the justifiability of the officer's actions: identifying more with Black compared to Blue Lives Matter meant people were more

likely to see the officer's actions as unjustified and, consequently, to interpret the harm as intentional (Figure 12).

Figure 12

Mediation Model Depicting the Pathway from Strength of Identifying with Black Compared to Blue Lives Matter to Perceived Intentions via Perceived Justifiability for Experiment 6

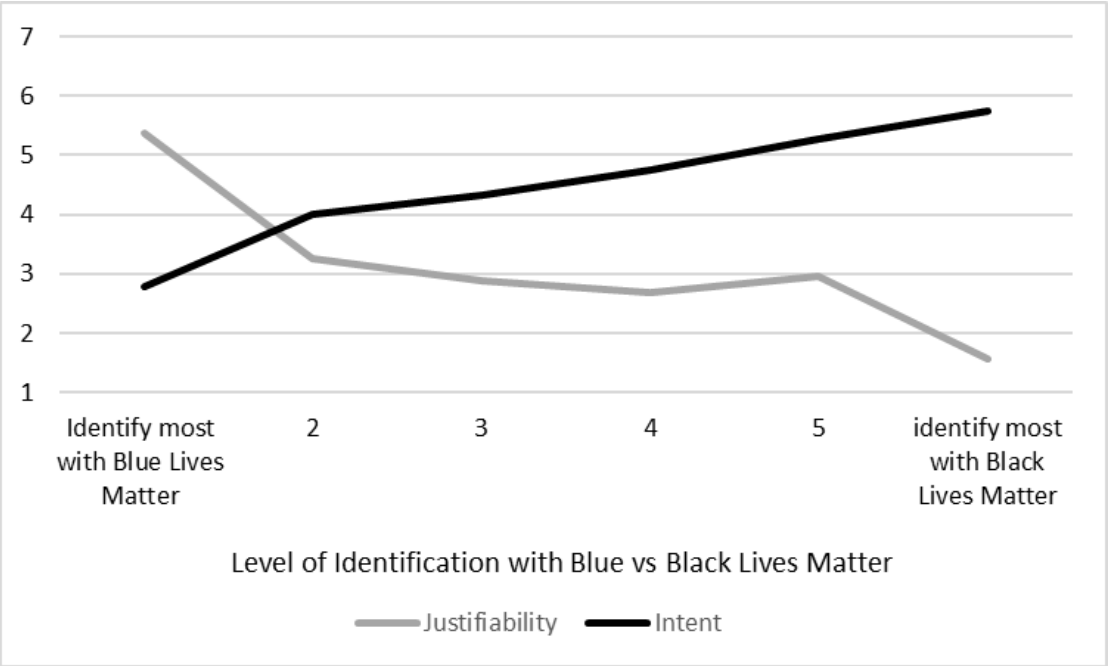


Exploratory Analyses

Polarized Extremes. Notably, a spike in judgements of justifiability and intentionality occurs at the extreme ranges of the identification scale (Figure 13). People at the extreme end of identifying most Blue Lives Matter judged the harm as far more justified (M_{DIFF} = 2.12, d = 1.13) and less intentional (M_{DIFF} = -1.12, d = 0.65) than varying one point off towards the middle (i.e., the mean difference in judgements is major when comparing between 5 and 6 on the scale). When moving off the extreme, however, the difference becomes more subtle (the mean difference in judgements is minor when comparing people on a 4 vs 5 on the scale than 5 vs 6) until you get to the opposite extreme of people who identified most with Black Lives Matter who judged the harm as far less justified (M_{DIFF} = -1.38, d = 0.82) and more intentional (M_{DIFF} = 0.48, d = .30) than varying one point off back towards the middle. Most of the conflicting opinions of whether the officer intentionally caused harm thus appears to be between the minority of people at either extreme (43%) than people at neither extreme (57%).

Figure 13

Mean Justifiability and intentionality responses in Experiment 6 by Level of Identification with Black vs Blue Lives Matter



Note. Responses to the measure of identification with Black vs Blue lives Matter was recorded on a 6-point Likert scale where 1 = identify most with Blue Lives Matter and 6 = identify most with Black Lives Matter. Intentionality and justifiability were measured on 7-point Likert scales where 1 = definitely not, and 7 = definitely yes.

Prior Knowledge. Overall, 64% of participants had seen the video prior to taking the study. Prior knowledge had no effect on any of the key variables (all *p*'s > .08). While prior knowledge had no effect, a greater percentage of Liberals (67%) compared to Conservatives (52%) and had seen the video before. Thus, results indicate a relationship between ideology and media consumption.

Exploring Base-rates. Consistent with TJM, paired with an overall tendency to perceive the harm as intentional (*M* = 4.97, *SD* = 1.85) was a tendency to judge the officer's actions as unjustified (*M* = 2.67, *SD* = 2.01) (Table 8). Consistent with prior findings,

participants also judged concrete actions (pushing someone) as more intentional (M = 6.10, SD = 1.30) compared to abstract outcomes (harming someone) (M = 4.97, SD = 1.85).

Table 8

Participant Responses in Experiment 6

	Intentionally Push	Intentionally Harm	Justified	Blame
Mean	6.10	4.97	2.67	5.68
Standard deviation	1.30	1.85	2.01	1.65

Note. Items were measured on separate 7-point Likert scales with higher numbers indicating higher perceived intentionality and justifiability, and more attributed blame.

Support for Blue Lives Matter was fairly high overall, but not as high as Black Lives Matter (see Table 9). Liberals tended to identify with Black more than Blue Lives Matter and to support Black Lives Matter more than Blue Lives Matter. Conservatives also tended to identify more with Black Lives Matter (but less so than Liberals).

Table 9

Level of Identifying with, and Support for, Black and Blue Lives Matter in Experiment 6

	Overall	Liberals	Conservatives
Identify Black vs Blue	4.48 (1.58)	5.09 (1.23)	3.70 (1.73)
Support Black	4.41 (1.72)	5.04 (1.39)	4.16 (1.35)
Support Blue	3.08 (1.72)	2.46 (1.65)	3.70 (1.96)

Note. level of identifying with Black compared to Blue Lives Matter was measured on a scale of 1-6 where 1 = more strongly identify with Blue Lives Matter and 6 = more strongly identify with Black Lives Matter. Participants were treated as Liberal/Conservative if they responded either “Very Liberal/Conservative”, “Liberal/Conservative”, “Somewhat Liberal/Conservative” to the prompt “I consider myself to be … On Overall Identification”.

Discussion

Experiment 6 showed that people’s political identities, like culture, can inform their judgments of justification and intentionality. Despite observing the same situation, peoples’ political beliefs affected whether they judged the harm to a protester as justified, and, consequently, whether the harm to the protestor was caused intentionally. People who

supported Black Lives Matter more than Blue Lives Matter perceived the harm to a protester as less justified, and, consequently, more intentional; whereas people who identified as supporting the Blue Lives Matter movement were more likely to view the officer's behavior as justified and the harm as unintentional.

Crucially, the effects in this experiment were much stronger than in Experiment 5: rather than one group/nationality judging the harm as *less* intentional, people at either ideological extreme made *opposite* judgements—one group judged it as *intentional and unjustified* whereas the other judged it as *not intentional and justified* (Figure 13). The flow on effects of this are radical. Perceived intentionality underlies legal decisions—was it murder or manslaughter?—political decisions—to cancel or to propagate someone?—and social decisions—is this person or group friend or foe? A jury of people where one ideology is dominant may see a harm as intentional and convict the assailant whereas a jury with a different dominant ideology may see it as unintentional and allow the assailant to go free, for example. While these implications may seem pessimistic to some, the mechanism can also provide hope—group conflict may be able to be reduced if people can come to see the justifying reasons behind harmful actions and thus that the harm may not have actually been intentional. However, these implications are purely speculative, future research needs to test the flow on effects.

Indeed, we are cautious of over claiming based on the results of one experiment. However, the results of the previously noted observer effects are consistent with our explanation, further supporting our claims. Moreover, previous research on TJM as well as Experiments 1-4 all indicate the validity of the proposed mechanism. Experiments 5 & 6 simply verify an important extension of this mechanism—people's perceptions of others are informed by their personal identities (which group membership is a part of).

General Discussion

Eight preregistered experiments found that people utilize information about the justifiability of causing a harm when determining whether it was caused intentionally. In particular, these experiments found that people judge less justified harms as more intentional across a variety of situations (Experiments 1-4), and that people update their judgements of intentionality when given new (un)justifying information (Experiment 4). These findings are all predicted by the Trade-off Justification Model (TJM), according to which causing harm without justification signals that the person harmed intentionally. Crucially, the last two experiments support an important extension of the hypothesized link between justifiability and intentionality—observer's cultural (Experiment 5) and group (Experiment 6) identities set the norms/values they use to evaluate the justifiability of causing harm and, thereby, the intentionality of different acts. People's perceptions of others are thus informed by their personal identities.

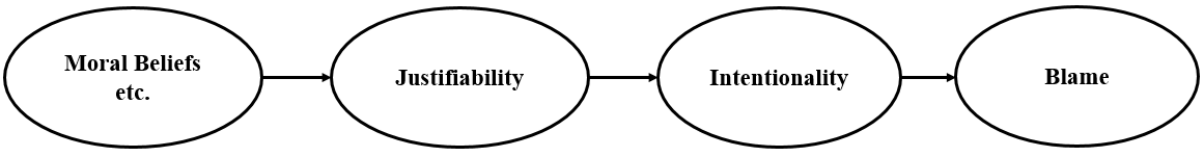
Research supporting theories of attributions based on social information have been used to support the idea that intention judgements can be explained without appealing to the observer's moral convictions (a part of people's identities)—information from the context is sufficient to explain people's judgements (Uttich & Lombrozo, 2010, p.99). An observer may use whether someone's actions violate a moral norm to inform their intention judgements, but their moral beliefs are not strictly part of this process. Rather, moral norms function as a particular piece of social information that can inform intentionality judgements. Our results show that the influence of observers' dispositional beliefs may be more prominent than previous social information models have indicated.

Figure 14 shows TJM's hypothesized pathway for the effect of morality on attributions of intentionality and blame: Observers' moral beliefs (among various other things) affect whether the harm is judged as justified. Judgements of whether the harm is

justified affect intentionality judgements (unjustified harms are judged as intentional). Harms that are intentional are judged as more blameworthy.

Figure 14

Hypothesized pathway for Morality's Effect on Attributions of Intentionality and Blame



That morality and blame are linked is not the issue. Rather, it is where in the pathway morality is exerting its influence that is problematic. Reflectively, morality should arguably come after intentionality: A judgement has been made that a person caused harm intentionally or unintentionally, *then* this is factored into judgements of whether they did something morally wrong, with the end result being a judgement of blameworthiness. That morality seems to be earlier in the process—before intentionality judgements—is the issue.

A modest take of our findings and this issue highlight a potential boundary condition for the rational use of social information: when there is a lack of information provided from the context about whether the harm is justified, people may use their own moral (and other) beliefs to judge the justifiability of the agent's actions and infer their intentions. However, when information about whether the harm was justified is provided, people use that. Thus, the issue (people's moral beliefs influencing intentionality judgements) only arises when the context is not explicitly providing information about the justifiability of the harm.

On this modest version of TJM people *may* be theorizing about or simulating the mental states of the other person to see whether the action was justified *for them* within the context to infer intentionality. It is only when observers do not have enough information that they cannot theorize properly about or simulate accurately the person's mental states that observers default to their own. Thus, when enough social information is provided, observer's own morals beliefs etc., may not be affecting their judgements of intentionality. Instead,

observers are using the moral beliefs of the agent as a particular kind of mental state that can affect behavior, much like people may for other types of mental states (beliefs, desires, etc.).

There is some evidence to support this modest version of TJM (e.g., Hindriks et al., 2016; Uttich & Lombrozo, 2010). For example, in a study by Uttich & Lombrozo (2010; Experiment 2) participants read about some supervillain's henchmen who were tasked with being the badest of the bad, "...never passing up a chance to spread malice and evil (p.91)". Analogous to the CEO example, a proposal is brought to one of the henchmen to rob a bank, which causes a harmful side-effect (increasing people's susceptibility to a poison the supervillain is distributing by using a neurotoxin to incapacitate people in the bank) to which the henchman replies that they do not care about the harmful side-effect, they just want to get as much money as they can. Contrary to the CEO example, however, people tended to say the harmful side-effect was not intentional. Thus, when the context provides salient reason to believe the harm was justified (in this case it was presumably justified because the explicit norm is to do as much harm as possible) people seem to be using this information, rather than their own beliefs, to judge intentions.

Uttich & Lomborozo's (2010) results support a modest version of TJM by showing that observers *can* put aside their own beliefs and use the information provided. Critically, however, their scenario is rooted in a comic book like world where the authors go at length to provide a very explicit norm which is removed from the often messy and constantly changing world of norms in which people reside. The agent being judged in the scenario (a henchman of a supervillain) is also far removed from the typical everyday person that people are making their attributions about. Thus, it is still uncertain whether in more messy, 'real' cases whether observers will adopt the information from the context or stick to their own beliefs. Future research should investigate the degree to which this happens outside of the lab with the cognitive and time constraints people can be under when making these judgements.

The modest version of TJM still distinguishes between judgements of justifiability made from a non-meta perspective (the observer's own) and those made from different meta-perspectives (from the agents or from the common point of view). The boundary condition provides some evidence that the perspective taken may change from a non-meta perspective when certain conditions are met (i.e., when there is enough information to take on the agent's or the social contexts perspective), though more research is required. We have some studies planned that will address that by testing the role of perspective-taking in judgments.

Alternative Models

It is no secret that psychology is inundated with models. The study of intentionality is no exception—even just a quick skim through the literature reveals a plethora of models each with their own unique take on what drives people's judgements (e.g., deep-self models, Sripada, 2010; norm violation models, Uttich & lombrozo, 2010; trade-off models, Machery, 2008; culpable control models, Alicke & Rose, 2010; see Cova, 2016; Knobe, 2010 for reviews). There are so many models that it can be hard for even a seasoned reader to tell the subtle differences between them and even harder to tease apart their predictions. Indeed, often different models are consistent with some of the same results, as is the case here. For example, motivated blame models (e.g., Alicke, 2000, 2008), could be consistent with the finding that people judge harms caused to ingroup members (e.g., people supporting the same political cause, Experiment 6) as more intentional because of an increased desire to blame those who cause ingroup members' harm. One could also argue that in Experiment 5 people of one nationality judged the harm as more intentional because in China is it the norm to send your children to cram school but not in America (Holton, 2010; Uttich & Lombrozo, 2010). However, TJM has the advantage of generating the predictions *a-priori*—the Experiments were designed specifically to test TJM, and all results were consistent with the model's

predictions. Moreover, alternate models may predict the same outcome but would not predict the same mediational pathway.

Crucially, there are some models the results distinctly contradict, for example Machery's (2008) Trade-off model. Machery (2008) argues that people infer intentionality from a person's willingness to incur costs in a tradeoff. For example, the CEO incurs the cost of harming the environment to gain the benefit of profit, so people think the harm was intentionally incurred. Rather than this, our results show that it is the perceived willingness of the agent to incur costs *without sufficient justification* that reveals their intentions—harm alone does not lead to intention judgments, harm *without justification* does.

Limitations

In many of our experiments (e.g., Experiment's 2a, b, 3a, b, 5) the harms were notably less severe than in previous research (e.g, Knobe, 2003). Lesser harms were chosen in Experiments 2a, b, 3a, and b so that increases in profit could affect the justifiability of the harm. If we used severe harms (e.g., environmental destruction, murder) then people may have perceived the harm as taboo and manipulating justifiability via increasing one factor (profit) would not be possible (Tetlock, 2003; Tetlock et al., 2000). As noted in Experiment 5's introduction, we chose a minor harm so that we could keep as much consistent across the vignettes as possible while still having the harm question make sense to people from both nationalities.

One key point of evidence for TJM comes from the consistent mediational pathway—as reported, the independent variable (e.g., profit) affects intention judgements via perceived justifiability. However, the mediator (perceived justifiability) was measured at the same time as the dependent variable (perceived intentionality) and thus cannot establish causality (Bullock et al., 2010). Nonetheless, the results are consistent with the proposed causal model.

Many experiments in the literature (see Knobe, 2010 for a review) have used a forced dichotomous yes/no response which does not capture the full range of judgements—sometimes people really cannot tell whether something was brought about intentionally. Measures thus need to include a point for people to indicate when they are uncertain what a person intended. Our scales included such a measure. However, this change complicates comparisons with previous results. We have used the end points of the scale as a clear indication that people thought the harm was/was not intentionally caused. It is unclear how well this transfers to dichotomous scales. For instance, it could be that people responding yes on a dichotomous scale would report anywhere from a 5-7 on ours, or that people without complete confidence that it was intentional default to responding 'no' on dichotomous scales.

Constraints on Generalizability

Children as young as 4-5 have been shown to judge unjustified harmful side-effects as intentional (Leslie et al., 2006; Pellizzoni et al., 2009). The pathway for judging intentionality we specify is cognitively complex, involving an understanding of the social context and even potentially shifting one's perspective. Given there is much cognitive development still to occur in young children and the complexity of the specified pathway, it is uncertain whether children are using the same pathway to arrive at their judgements. Future research should test the effects of manipulating the justifiability of harm on children's judgements.

In order to acquire large enough samples to sufficiently power our studies we used online crowd-sourcing platforms (i.e., Mechanical Turk, Prolific) and University student pools. Using multiple sources of recruitment allows generalization between more demographics than student pools alone. However, many demographics were not tested. For example, people with deficits in social cognition/theory of mind (e.g., people who have autism or Asperger's) were not tested and may be less likely to show this effect due to their

limited ability to perceive and incorporate social information. For years psychological findings have been limited by being tested on WEIRD populations (Henrich, et al., 2010). New crowdsourcing platforms such as Mechanical Turk and Prolific have allowed us to collect data from 6 different countries that vary in cultural distance (see https://world.culturalytics.com for the exact cultural distance values; Muthukrishna et al., 2020). We found support for the same pathway in each country, indicating that our model is culturally generalizable.

Implications and Extensions of Research

Judgements of intentionality have consequences—they impact legal, political, and social decisions, for example. Here, we successfully tested and extended a model for how people decide whether a harm was intentional—unjustified harms are judged intentional. This is important because what counts as unjustified varies because of individual differences in values and culture, thus potentially creating misunderstandings of whether a harm was intentional. Our research therefore suggests some potentially problematic ways that these highly consequential judgements might differ.

For instance, laws are conceived and derived from values (Allsop, 2017) but also set the values or standards from which people within the legal system (e.g., judges, jury members) are required to make their judgements from. Different legal systems based on different values could essentially provide different perspectives from which to judge the justifiability of a harm and thus may lead to different convictions despite the facts of the case being exactly the same. For example, people within a legal system based on strong values of honor may perceive harms caused to a person's reputation as far less justified and more intentional than people within a legal system that values honor comparatively less, leading to different sentences. Future research should investigate how the expression of different values

across legal systems (such as honor or justice) may affect the perceived justifiability and intentionality of harms. If the different values across legal systems affect the perceived justifiability and intentionality of harms, this raises serious questions about how judges and juries should establish intentionality for defendants whose values differ from the system in which they are being trialed.

Problems caused by judgements of justifiability based on different norms/values also likely occur outside of legal contexts. Indeed, this difference in values may underlie some of the conflicts between different political groups—people may be seen as having malicious, harmful intentions for doing things that are justified by their group's standards. Martin Gugino's case is an example of this. Most people who identified strongly with Blue Lives Matter saw the officer as using a justified means to disperse a protestor breaking curfew, and thus perceived no harmful intentions. Most people who identified strongly with Black Lives Matter, however, saw the officer as unjustifiably putting down a legitimate protester, and perceived the harm as clearly intentional. One group thus had reason to support the officer's suspension, whereas others saw the reason behind the officer's suspension as unjustified and protested it (Li & Imam, 2020, June 6). Future research should test the extent to which people change their judgements upon learning a group's justification for causing the harm.

Concluding statement

Intentionality judgements are interesting, important, and somewhat mysterious too: the minds of others are invisible, so how are people determining whether others caused harm intentionally? Our research shows that people use information about the justifiability of causing the harm when judging whether it was intentional—harms caused without sufficient reasons are judged as intentional. Crucially, however, people's cultural and group identities can change the norms/values used to judge whether the harm was justified, and,

consequently, whether the harm is judged intentional. People's perceptions of others are thus informed by their personal identities, meaning that people's judgements of whether a harm was caused on purpose will cohere or conflict based on whether they agree or disagree that the harm was justified.

References

Ahlers, G. K. C., Cumming, D., Günther, C., & Schweizer, D. (2015). Signaling in Equity Crowdfunding. *Entrepreneurship Theory and Practice*, 39(4), 955–980. https://doi.org/10.1111/etap.12157

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences*, 33(4), 330-1. http://dx.doi.org/10.1017/S0140525X10001664

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

Alicke, M. (2008). Blaming Badly, *Journal of Cognition and Culture*, 8(1-2), 179-186. doi: https://doi.org/10.1163/156770908X289279

Allsop, J. "Values in law: how they influence and shape rules and the application of law." *Brief* 44, no. 2 (2017): 49-54.

Aquinas, T (13th c). Summa Theologica II-II, Q. 64, art. 7, "Of Killing", in *On Law, Morality, and Politics* (W. Baumgarth & R. Regan, Eds.). Hackett Publishing Co., 1988

Bahník, S., Vranka. M., & Trefná. K. (2021) What makes euthanasia justifiable? The role of symptoms' characteristics and interindividual differences, *Death Studies*, 45:3, 226-237, DOI: 10.1080/07481187.2019.1626945

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, 25(4), 474-498. https://doi.org/10.1111/j.1468-0017.2010.01398.x

Beebe, J. R., & Jensen, M. (2012). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical psychology*, 25(5), 689-715. https://doi.org/10.1080/09515089.2011.622439

Bird, R., & Smith, E. A. (2005). Signaling theory, strategic interaction, and symbolic capital. *Current Anthropology*, 46, 221–248.

Backwell, P. R., Christy, J. H., Telford, S. R., Jennions, M. D., & Passmore, J. (2000). Dishonest signalling in a fiddler crab. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1444), 719-724. https://doi.org/10.1098/rspb.2000.1062

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550–558. https://doi.org/10.1037/a0018933

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. *Infant behavior and development*, 21(2), 315-330. https://doi.org/10.1016/S0163-6383(98)90009-1

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 464-473. https://doi.org/10.1177/1948550619875149

Cokely, E. T., & Feltz, A. (2009). Adaptive variation in judgment and philosophical intuition. *Consciousness and Cognition*, 18(1), 356-358. https://doi.org/10.1016/j.concog.2009.01.001

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of management*, 37(1), 39-67. https://doi.org/10.1177/0149206310388419

Cova, F., Lantian, A., & Boudesseul, J. (2016). Can the Knobe effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality. *Personality and Social Psychology Bulletin*, 42(10), 1295-1308. https://doi.org/10.1177/0146167216656356

Cova, F (2016). The Folk Concept of Intentional Action: Empirical approaches. In Wesley Buckwalter & Justin Sytsma (eds.), Blackwell Companion to Experimental Philosophy.

Cushman, F., & Mele, A. (2008). Two-and-a-half folk concepts? In J.Knobe & S.Nichols (Eds.), Experimental philosophy (pp. 171–188). Oxford, England: Oxford University Press.

Detweiler, R. A. (1975). On inferring the intentions of a person from another culture. *Journal of Personality*, 43(4), 591-611. https://doi.org/10.1111/j.1467-6494.1975.tb00724.x

Faul, F., Erdfelder, E., Buchner, A. et al. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 1149–1160 (2009). https://doi.org/10.3758/BRM.41.4.1149

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological science*, 18(3), 233-239. https://doi.org/10.1111/j.1467-9280.2007.01882.x

Gilbert, D. T. (1998). Ordinary personology. In Gilbert, D., Fiske, S. T., & Lindzey, G. (4th ed) The handbook of social psychology, 2, 89-150.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2), 145-171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4), 517.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? resolving a controversy over intentionality and morality. *Personality & Social Psychology Bulletin*, 36(12), 1635-1647. https://doi.org/10.1177/0146167210386733

Guglielmo, S. (2015). Moral judgment as information processing: An integrative review. *Frontiers in Psychology*, 6, 1637. https://doi.org/10.3389/fpsyg.2015.01637

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2020. World Values Survey: Round Seven - Country-Pooled Datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat.

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality & Social Psychology Bulletin*, 32(10), 1402-1413. https://doi.org/10.1177/0146167206291006

Heider, F. (1958). The psychology of interpersonal relations. Lawrence Erlbaum. https://doi.org/10.4324/9780203781159

Henrich, J., Heine, S. J., Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hindriks, F., Douven, I., & Singmann, H. (2016). A new angle on the knobe effect: Intentionality correlates with blame, not with praise: A new angle on the knobe effect. *Mind & Language*, 31(2), 204-220. https://doi.org/10.1111/mila.12101

Hindriks, F. (2014). Normativity in action: How to explain the Knobe effect and its relatives. *Mind & Language*, 29(1), 51–72. https://doi.org/10.1111/mila.12041

Holton, R. (2010). Norms and the knobe effect. *Analysis (Oxford)*,70(3), 417-424. https://doi.org/10.1093/analys/anq037

Hume. D. (1740) A Treatise of Human Nature. In Norton. D and Norton. M (eds.), Oxford, Clarendon Press, 2007.

Jiang, D., Li, T., & Hamamura, T. (2015). Societies' tightness moderates age differences in perceived justifiability of morally debatable behaviors. *European Journal of Ageing*, 12(4), 333-340. https://doi.org/10.1007/s10433-015-0346-z

Jones, E. E., Davis, K. E. (1965). From acts to dispositions the attribution process in person perception. In Berkowitz, L. (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). Academic Press.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the cause of behavior. In E. E.Jones, D. E.Kanouse, H. H.Kelley, R. E.Nisbett, S.Valins, & B.Weiner (Eds.), Attribution: Perceiving the causes of behavior (pp. 79–94). Morristown, NJ: General Learning Press.

Karasek III, R., & Bryant, P. (2012). Signaling theory: Past, present, and future. *Academy of Strategic Management Journal*, 11(1), 91.

Knittle, A (2017, December 17) Accused killer Jerrod Murray was feared by Oklahoma college dorm neighbors. *The Oklahoman*. https://www.oklahoman.com/article/3737992/accused-killer-jerrod-murray-was-feared-by-oklahoma-college-dorm-neighbors

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63(3), 190–194.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33(4), 315–329. https://doi.org/10.1017/S0140525X10000907

Larmer, B (2014, December 31) Inside a Chinese Test-Prep Factory. *The New York Times Magazine.* https://www.nytimes.com/2015/01/04/magazine/inside-a-chinese-test-prep-factory.html

Laurent, S. M., Clark, B. A. M., & Schweitzer, K. A. (2015). Why side-effect outcomes do not affect intuitions about intentional actions: Properly shifting the focus from

intentional outcomes back to intentional actions. *Journal of Personality and Social Psychology*, 108(1), 18-36. https://doi.org/10.1037/pspa0000011

Leslie, A. M., Knobe, J., Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17(5), 421–427. https://doi.org/10.1111/j.1467-9280.2006.01722.x

Li, D & Imam, J (2020, June 6). Buffalo police officers resign from unit in protest of suspended colleagues who shoved man, 75, to ground. *NBC News*. https://www.nbcnews.com/news/us-news/buffalo-officers-shoving-75-year-old-ground-decried-governor-where-n1225776

Li, J., & Tomasello, M. (2018). The development of intention-based sociomoral judgment and distribution behavior from a third-party stance. *Journal of Experimental Child Psychology*, 167, 78-92. https://doi.org/10.1016/j.jecp.2017.09.021

Li, J., Hou, W., Zhu, L., & Tomasello, M. (2020). The development of intent-based moral judgment and moral behavior in the context of indirect reciprocity: A cross-cultural study. *International Journal of Behavioral Development*, 44(6), 525-533. https://doi.org/10.1177/0165025420935636

Liao, Y., Sun, Y., Li, H., Deák, G. O., Feng, W. (2018). Intensity of caring about an action's side-effect mediates attributions of actor's intentions. *Frontiers in Psychology*, 9, 1329. https://doi.org/10.3389/fpsyg.2018.01329;

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). 12- and 18-month-olds point to provide information for others. *Journal of Cognition and Development*, 7(2), 173-187. https://doi.org/10.1207/s15327647jcd0702_2

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, 23(2), 165–189. https://doi.org/10.1111/j.1468-0017.2007.00336.x

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33(2), 101-121. https://doi.org/10.1006/jesp.1996.1314

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895. https://doi.org/10.1037/0033-2909.132.6.895

McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology*, 22(2), 171-193. https://doi.org/10.1037/h0032602

Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31(1), 184-201. https://doi.org/10.1111/j.1475-4975.2007.00147.x

Moll, H., & Tomasello, M. (2007). How 14-and 18-month-olds know what others have experienced. *Developmental psychology*, 43(2), 309. https://doi.org/10.1037/0012-1649.43.2.309

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, 146(1), 123. https://doi.org/10.1037/xge0000234

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*, 116(2), 215–236. https://doi.org/10.1037/pspa0000137

Monroe, A. E., Reeder, G. D., & James, L. (2015). Perceptions of intentionality for goal-related action: Behavioral description matters. *PloS one*, 10(3), e0119841. https://doi.org/10.1371/journal.pone.0119841

Moss, T. W., Neubaum, D. O., & Meyskens, M. (2015). The effect of virtuous and entrepreneurial orientations on microfinance lending and repayment: A signaling

theory perspective. *Entrepreneurship theory and practice*, 39(1), 27-52. https://doi.org/10.1111/etap.12110

Murphy, R.H. The rationality of literal Tide Pod consumption. *J Bioecon* 21, 111–122 (2019). https://doi.org/10.1007/s10818-019-09285-1

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, 31(6), 678-701. https://doi.org/10.1177/0956797620916782

Nadelhoffer, T. (2004). Blame, badness, and intentional action: A reply to Knobe and Mendlow. https://doi.org/10.1037/h0091247

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 9(2), 203–219. https://doi.org/10.1080/13869790600641905

Nakamura, K. (2018). Harming is more intentional than helping because it is more probable: The underlying influence of probability on the knobe effect. *Journal of Cognitive Psychology*, 30(2), 129–137. https://doi.org/10.1080/20445911.2017.1415345

Ogunfowora, B., Stackhouse, M. & Oh, WY. Media Depictions of CEO Ethics and Stakeholder Support of CSR Initiatives: The Mediating Roles of CSR Motive Attributions and Cynicism. *J Bus Ethics* 150, 525–540 (2018). https://doi.org/10.1007/s10551-016-3173-z

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior*, 30(2), 114-123. https://doi.org/10.1016/j.evolhumbehav.2008.09.004

Park, J., Chae, H., & Choi, J. N. (2017). The need for status as a hidden motive of knowledge-sharing behavior: An application of costly signaling theory. *Human Performance*, 30(1), 21–37. https://doi.org/10.1080/08959285.2016.1263636

Pellizzoni, S., Siegal, M., & Surian, L. (2009). Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology*, 45(1), 289–295. https://doi.org/10.1037/a0014165

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & language*, 24(5), 586-604. https://doi.org/10.1111/j.1468-0017.2009.01375.x

Phelan, M.T., Sarkissian, H. The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138, 291–298 (2008). https://doi.org/10.1007/s11098-006-9047-y

Phelan, M., Sarkissian, H. (2009). Is the "Trade-off Hypothesis" worth trading for? *Mind & Language*, 24(2), 164–180. https://doi.org/10.1111/j.1468-0017.2008.01358.x

Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), The social psychology of morality: Exploring the causes of good and evil (pp. 91–108). American Psychological Association. https://doi.org/10.1037/13091-005

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self Versus Others. *Psychological Review*, 111(3), 781–799. https://doi.org/10.1037/0033-295X.111.3.781

Robbins, E., Shepard, J., & Rochat, P. (2017). Variations in judgments of intentional action and moral evaluation across eight cultures. *Cognition*, 164, 22-30. https://doi.org/10.1016/j.cognition.2017.02.012

Rose, J & Levenson, E (2020, June 16). Buffalo protester Martin Gugino has a fractured skull and cannot walk. *CNN*. https://edition.cnn.com/2020/06/16/us/martin-gugino-protester-skull/index.html

Rose, J Kim, A & Johnson, E (2020, June 30). Buffalo protester Martin Gugino has been released from the hospital. *CNN*. https://edition.cnn.com/2020/06/30/us/martin-gugino-released-hospital-trnd/index.html

Rowe, S. J., Vonasch, A. J., & Turp, M.-J. (2020). Unjustifiably Irresponsible: The Effects of Social Roles on Attributions of Intent. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550620971086

Sauer, H., Bates, T. (2013). Chairmen, cocaine, and car crashes: The Knobe effect as an attribution error. *The Journal of Ethics*, 17(4), 305–330. https://doi.org/10.1007/s10892-013-9150-1

Sauer, H. (2014). It's the Knobe effect, stupid!. *Review of Philosophy and Psychology*, 5(4), 485-503. https://doi.org/10.1007/s13164-014-0189-0

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality*, 47(5), 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation Through Attribution of False Belief by 2-Year-Olds. *Psychological Science*, 18(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Spaulding, S. (2018). How we understand others: Philosophy and social cognition. Routledge.

Spence, M. (1978). Job market signaling. In Uncertainty in economics (pp. 281-306). Academic Press. https://doi.org/10.1016/B978-0-12-214850-7.50025-5

Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, 151(2), 159-176. https://doi.org/10.1007/s11098-009-9423-5

Tannenbaum, D., Ditto, P. H., Pizarro, D. A. (2007). Different moral values produce different judgments of intentional action [Unpublished manuscript]. University of California–Irvine

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78(5), 853–870. https://doi.org/10.1037/0022-3514.78.5.853

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in cognitive sciences*, 7(7), 320-324. https://doi.org/10.1016/S1364-6613(03)00135-9

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, 10(1), 72–81. https://doi.org/10.1177/1745691614556679

Uttich, K., Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1), 87–100. https://doi.org/10.1016/j.cognition.2010.04.003

Van Vugt, M., & Hardy, C. L. (2010). Cooperation for reputation: Wasteful contributions as costly signals in public goods. *Group Processes & Intergroup Relations*, 13(1), 101–111. https://doi.org/10.1177/1368430209342258

Voiklis, J., & Nickerson, J. V. (2012). Tort Reform: Cognitive Perspective Taking Promotes Attributions of 'Oblique' Intent for Side-Effects of Intentional Action. *Available at SSRN 2144205*. http://dx.doi.org/10.2139/ssrn.2144205

Vonasch, A. J., Baumeister, R. F. (2017). Unjustified side effects were strongly intended: Taboo tradeoffs and the side-effect effect. *Journal of Experimental Social Psychology*, 68, 83–92. https://doi.org/10.1016/j.jesp.2016.05.006

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of theoretical Biology*, 53(1), 205-214. https://doi.org/10.1016/0022-5193(75)90111-3

Appendices

Appendix A: Vignettes and Questions in Experiment 1

1. Sally saw a runaway trolley barreling down the railway tracks. Ahead, on the tracks she saw 5 people tied up and unable to move with the trolley heading straight for them. She saw a lever that could switch the trolley to a different set of tracks where only one person was tied up and unable to move. Sally pulled the lever, saving the 5 people but killing the other.
   Did Sally intentionally kill the other person?
   Was Sally's decision to pull the lever justified?

2. Mark saw a runaway trolley barreling down the railway tracks. Ahead on the tracks he saw his mother tied up and unable to move with the trolley heading straight for her. He saw a lever that could switch the trolley to a different set of tracks where a stranger was tied up and unable to move. Mark pulled the lever, saving his mother but killing the stranger.
   Did Mark intentionally kill the stranger?
   Was Mark's decision to pull the lever justified?

3. Jim, a prisoner of war, was given a choice by his captors. In front of Jim there was a row of 20 other prisoners of war tied up against a wall. Jim could choose to shoot one and the other 19 would be set free, or he could not shoot anyone, and all of them would be shot. Jim decided to shoot one of the prisoners.
   Did Jim intentionally kill the prisoner?
   Was Jim's decision to shoot the prisoner justified?

4. Jacob threw a party to be more popular. Jacob knew his party would make Curtis, his roommate, fail the morning's exam.
   Did Jacob intentionally make Curtis fail the morning's exam?
   Was Jacob's decision to throw a party justified?

5. Russell planted a tree to decorate his yard. Russell knew his tree would block his neighbor's sun, making them unhappy.
   Did Russell intentionally make his neighbor unhappy?
   Was Russell's decision to plant the tree justified?

6. A Party was thinking about implementing a new policy. The policy would allow them to remain politically popular by increasing GDP but would also increase deaths. The party decided to implement the policy. Sure enough, the amount of death increased.
   Did the Party intentionally increase deaths?
   Was the party's decision to implement the policy justified?

7. Eugene screamed during a tennis match to express his excitement, but Eugene knew his screaming would also put his opponent off.

Was Eugene intentionally putting off his opponent?
Is Eugene justified in screaming during the match?

8. Curtis released the documents to gain publicity. Curtis knew the documents would ruin his friend's reputation.
   Did Curtis intentionally ruin his friend's reputation?
   Was Curtis' decision to release the documents justified?

9. The abolitionist movement in Laputa was aimed at ending slavery in the country. However, supporters of this movement also knew that ending slavery would have negative effects on the economy. The movement was successful and, sure enough, there were negative effects on the country's economy.
   Did the supporters intentionally harm the economy?
   Were the supporters justified in ending slavery?

10. Affirmative action is aimed at increasing the opportunities provided to underrepresented parts of society. This movement will increase opportunity for underrepresented groups but also decrease opportunity for overrepresented groups.
    Is this movement intentionally decreasing opportunity for overrepresented groups?
    Is the movement justified in decreasing opportunity for overrepresented groups?

11. Mary, a doctor, was thinking about going on a spontaneous holiday. She knew if she cancelled all her appointments and left that not all her patients would be able to rebook with another doctor, causing them to become more ill than they otherwise would have. Mary decided to cancel all her appointments and go on holiday anyway. Sure enough, not all of her patients were able to rebook with another doctor, causing some of them to become more ill than they otherwise would have.
    Did Mary intentionally cause her patients to become more ill than they otherwise would have?
    Was Mary's decision to go on holiday justified?

12. Jane, a doctor, got a call that her dog was missing. She knew if she cancelled all her appointments to go find her dog that not all of her patients would be able to rebook with another doctor, causing them to become more ill than they otherwise would have. Jane decided to cancel all her appointments and go find her dog anyway. Sure enough, not all of her patients were able to rebook with another doctor, causing some of them to become more ill than they otherwise would have.
    Did Jane intentionally cause her patients to become more ill than they otherwise would have?
    Was Jane's decision to go find her dog justified?

13. The president of a developing country knew about a coal mining programme that would bring the country great economic prosperity but also greatly harm the environment. The president implemented the programme. Sure enough, the programme brought the country great economic prosperity but also greatly harmed the environment.
    Did the president intentionally harm the environment?

Was the president's decision to implement the program justified?

14. The president of a developed country knew about a coal mining programme that would bring minor benefits to the economy but also greatly harm the environment. The president implemented the programme. Sure enough, the programme brought minor benefits to the economy but also greatly harmed the environment.
    Did the president intentionally harm the environment?
    Was the president's decision to implement the program justified?

15. Jo owns a donut store. He was thinking of closing the store early so he could go home and look after his very sick child, but Jo knew that closing early would annoy the customers who were coming to buy his donuts. Jo decided to close anyway. Sure enough, the customers arrived and the place was closed, really annoying them.
    Did Jo intentionally annoy the people?
    Was Jo's decision to close his store justified?

16. Mark is a mechanic who owns a garage.He was thinking of closing the store early so he could go home and get ready for his date, but Mark knew that closing early would annoy Fred who had an appointment that evening to get his car fixed. Mark decided to close anyway. Sure enough, when Fred arrived to get his car fixed he found the garage closed, really annoying him.
    Did Mark intentionally annoy Fred?
    Was Fred's decision to close his garage justified?

17. Josh is an on call emergency worker. He turned his buzzer off so that he would not get disrupted on his date. A person called in needing urgent medical help but Josh did not pick up. Luckily, they were able to get someone else to help.
    Did Josh intentionally put people at risk by turning his buzzer off?
    Was Josh's decision to turn off his buzzer justified?

18. Frank is an on call emergency worker. He went to go visit his dying wife in hospital and turned his buzzer off so that their last moments together would not be disrupted. A person called in needing urgent medical help but Frank did not pick up. Luckily, they were able to get someone else to help.
    Did Frank intentionally put people at risk by turning his buzzer off?
    Was Frank's decision to turn off his buzzer justified?

19. The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment." The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was harmed.
    Did the chairman of the board intentionally harm the environment?
    Was the chairman of the board's decision to run the program justified?

Appendix B: Vignettes and Questions in Experiment 4

New justifying information condition.

1. Ted

Ted answered a phone call during a movie, despite knowing it would annoy the other people in the theatre.

Did Ted intentionally annoy the other people in the theatre?
Was Ted justified in answering the phone call?

New info: Ted's mother was in hospital having open heart surgery and he'd been expecting a call from the hospital.

2. Frank

Frank is an on call emergency worker. He had turned his buzzer off for a moment so that he wouldn't be disrupted. A person called in needing urgent medical help but Frank did not pick up. Luckily, they were able to get someone else to help.

Did Frank intentionally put people at risk by turning his buzzer off?
Was Frank's decision to turn off his buzzer justified?

New info: Frank had gone to go visit his dying wife in hospital and turned his buzzer off so that their last moments together would not be disrupted.

3. CEO

The CEO of a company started a new programme to increase company profits despite knowing it would harm the environment.

Did the CEO intentionally harm the environment?
Was the CEO justified in starting the programme?

Update: The CEO's daughter had been kidnapped and they needed the company to make more money in order to be able to pay for her ransom.

4. Andrew

A little old lady came onto the bus and needed a place to sit. Andrew knew she'd stumble and hurt herself as the bus took off if she didn't get a seat. Yet, Andrew did not offer his seat to her. She didn't get a seat and, sure enough, as the bus took off the lady stumbled and hurt herself.

Did Andrew intentionally cause the old lady to get hurt?
Was Andrew justified in not giving up his seat?

New info:
Andrew had recently broken his knee and standing on the bus would have caused him great pain.

5. Jerry

Jerry and a friend are at a bar and they both leave the table briefly, leaving their jackets on the seats, to get some food and drinks. When they return, two female strangers are sitting in their seats. Jerry explains to them that they have been sitting there. Much to Jerry's annoyance they reply ''you'll have go find another table, and hand them their jackets back.''

Did the two women intentionally annoy Jerry?
Were the two women justified in taking the seat?

New info: The bar is extremely popular and in order to turn people in and out quickly has a very clear rule that no seats can be saved—you move you lose.

New justifying information condition.

6. Tom

While playing a round of golf with his dad, Tom tried a shot from an odd angle and ended up smashing one of the neighbouring houses windows.

Did Tom intentionally smash the neighbours window?
Was Tom Justified in trying a shot from an odd angle?

New info: Before taking the shot Tom's dad had yelled at him and warned him not to hit the ball that way as it would likely hit the neighbours window.

7. Jordan

A group of coworkers go out for lunch but do not invite Jordan, who was absent from her desk. Jordan was upset that the others did not invite her.

Are the coworkers intentionally upsetting Jordan?
Are the coworkers justified in going out for lunch without Jordan?

New Info: When they got back Jordan asked why they didn't invite her. One of the coworkers laughed and said that they didn't want to be associated with someone of her kind.

8. Angela

Angela filled her desk up with houseplants, making her desk smell lovely but really annoying one of her coworkers.

Did Angela intentionally annoy her co-worker?
Was Angela justified in bringing in the houseplants?

New info: Angela had been asked, many, many times by her co-worker to please not bring in any houseplants as it gave her bad allergies.

9. Gary

Gary used some of the communal supplies in his office while making a paper mache for his friend's birthday, annoying one of his colleagues.

Did Gary intentionally annoy his colleague?

Was Gary justified in taking the office supplies?

New info: Gary already had plenty of his own supplies to use but would still repeatedly take stuff from the communal supplies, despite knowing his coworker needed the supplies to complete an important work order.

10. Stephen
Stephen returned his students' graded assignments late, despite knowing this would annoy them.

Did Stephen intentionally annoy his students?
Was Stephen justified in handing the assignments back late?

New info: When another professor asked Stephen why he did not finish grading the assignments on time despite knowing it would annoy his students he replied, "I don't care that it annoyed them, I was busy finishing my own manuscript"

Irrelevant new information condition.

11. Sarah
Sarah snuck into her sister's room when she was out and took one of her tops to wear, despite knowing it would annoy her. Sure enough, when her sister found out, she was very annoyed.

Did Sarah intentionally annoy her sister?
Was Sarah justified in taking her sister's top?

New info: The top Sarah took was red.

12. Lisa
Lisa left her dirty dishes sitting on the bench all day despite knowing it would annoy her flatmates.

Did Lisa intentionally annoy her flatmates?
Was Lisa justified in leaving the dishes out?

New info: Lisa's birthday is on a Friday this year.

13. Paul
Paul's mother's birthday was coming up and a big celebration was being planned. Paul's sister had planned the event and given all the siblings strict instructions to be on time because of how sad their mother gets whenever someone's late to a family event—their mother believes that being late is a sign that a person doesn't want to be there. Despite all this forewarning, Paul still showed up late, making his mother sad.

Did Paul intentionally make his mother sad?
Was Paul justified in turning up late?

New info: Paul is the tallest out of all his siblings.

14. Jacob
Jacob threw a party despite knowing that it would make Curtis, his roommate, fail the morning's exam.

Did Jacob intentionally make Curtis fail the morning's exam?
Was Jacob's decision to throw a party justified?

New info: Jacob and Curtis both have hazel eyes.

15. Mike
Mike's coworker had been talking on the telephone for over an hour and a half despite knowing they were supposed to be helping Mike. Consequently, Mike had to stay late to get the project done before the deadline.

Did Mike's coworker intentionally cause Mike to have to stay late?
Was Mike's coworker justified in talking to their friend on the phone for over an hour and a half?

New info: Mike and his coworker are about the same height.